B191

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
C12N 15/12, 15/00, 15/11, 15/63, A61K 38/16, C07K 16/00, C12P 21/02, C12Q 1/68, G01N 33/53, 33/68

A1

(11) International Publication Number: WO 99/18208

(43) International Publication Date: 15 April 1999 (15.04.99)

(21) International Application Number: PCT/US98/20775

(22) International Filing Date: 1 October 1998 (01.10.98)

(30) Priority Data:

60/060,837	2 October 1997 (02.10.97)	US
60/060,862	2 October 1997 (02.10.97)	US
60/060,839	2 October 1997 (02.10.97)	US
60/060,866	2 October 1997 (02.10.97)	US
60/060,843	2 October 1997 (02.10.97)	US
60/060,836	2 October 1997 (02.10.97)	US
60/060,838	2 October 1997 (02.10.97)	US
60/060,874	2 October 1997 (02.10.97)	US
60/060,833	2 October 1997 (02.10.97)	US
60/060,884	2 October 1997 (02.10.97)	US
60/060,880	2 October 1997 (02.10.97)	US

(71) Applicant *(for all designated States except US)*: HUMAN GENOME SCIENCES, INC. [US/US]; 9410 Key West Avenue, Rockville, MD 20850 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: DUAN, D., Roxanne [US/US]; 5515 Northfield Road, Bethesda, MD 20817 (US). FLORENCE, Kimberly, A. [US/US]; 12805 Atlantic Avenue, Rockville, MD 20851 (US). ROSEN, Craig, A. [US/US]; 22400 Rolling Hill Road, Laytonsville, MD 20882 (US). RUBEN, Steven, M. [US/US]; 18528 Heritage Hills Drive, Olney, MD 20832 (US). GREENE, John, M. [US/US]; 872 Diamond Drive, Gaithersburg, MD 20878 (US). YOUNG, Paul [US/US]; 122 Beckwith Street, Gaithersburg, MD 20878 (US). FERRIE, Ann, M. [US/US]; 120 Fox Run Drive, Tewsbury, MA 01876 (US). YU, Guo-Liang [CN/US]; 1714C Marina Court, San Mateo, CA 94403 (US). JANAT, Fouad [US/US]; 140 High Street #202, Westerly, RI 02891 (US). NI, Jian [CN/US]; 5502 Manorfield, Rockville, MD 20853 (US). CARTER, Kenneth, C. [US/US]; 11601 Brandy Hill Lane, North Potomac, MD 20878 (US). ENDRESS, Gregory, A. [US/US]; 9729 Clagett Farm Drive, Potomac, MD 20854 (US). FENG, Ping [CN/US]; 4 Relda Court, Gaithersburg, MD 20878 (US). LAFLEUR, David, W. [US/US]; 3142 Quesada Street, N.W., Washington, DC 20015 (US). SHI, Yanggu [CN/US]; Apartment 102, 437 West Side Drive, Gaithersburg, MD 20878 (US).

(74) Agents: BROOKES, A., Anders et al.; Human Genome Sciences, Inc., 9410 Key West Avenue, Rockville, MD 20850 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, GM, HR, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published

With international search report.

With an indication in relation to deposited biological material furnished under Rule 13bis separately from the description.

(54) Title: 101 HUMAN SECRETED PROTEINS

(57) Abstract

The present invention relates to novel human secreted proteins and isolated nucleic acids containing the coding regions of the genes encoding such proteins. Also provided are vectors, host cells, antibodies, and recombinant methods for producing human secreted proteins. The invention further relates to diagnostic and therapeutic methods useful for diagnosis and treating disorders related to these novel human secreted proteins.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria
AU	Australia
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CS	Czechoslovakia
CZ	Czech Republic
DE	Germany
DK	Denmark
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgystan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakhstan
LI	Liechtenstein
LK	Sri Lanka
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SI	Slovenia
SK	Slovakia
SN	Senegal
TD	Chad
TG	Togo
TJ	Tajikistan
TT	Trinidad and Tobago
UA	Ukraine
US	United States of America
UZ	Uzbekistan
VN	Viet Nam

Description

A Method For Predicting Protein Structure

1. Introduction

The present invention relates to methods of predicting the tendency of a portion of a protein to form amphiphilic α or β structure.

2. Background Of The Invention

2.1. Methods For Determining Protein Structure

Several algorithms are currently used to evaluate the secondary structure of proteins, including the Kyte-Doolittle, Chou-Fasman-Prevelige, and PHD methods.

The Kyte-Doolittle method (Kyte and Doolittle, 1982, J. Mol. Biol. 157: 105-132) evaluates the hydrophobicity and hydrophilicity of each amino acid, as they appear sequentially in a protein. The program then uses a continuous moving segment approach that determines the average hydropathy within a predetermined segment. Although the program can accurately predict interior and exterior regions of soluble globular proteins, data on membrane spanning regions of transmembrane proteins is more ambiguous.

The Chou-Fasman-Prevelige (CFP) algorithm (Prevelige and Fasman, 1989, in "Predictions of Protein Structure and the Principles of Protein Conformation", Fasman, ed., Plenum Press, New York, pp. 391-416) uses a statistical approach to the study of protein secondary structur . The conformational parameters for each amino acid are calculated using the relative frequency of a given amino acid within a protein, its occurence in a given type of secondary structur , and the fraction of residues occuring in that type of

structure. Since these parameters (such as hydrophobicity) contain information about protein stability, properly weighted for their relative importance, they are useful for predicting secondary structures. These parameters, represented by Pα and Pβ or Pc (for α-helix, β-sheets or coils, respectively) are utilized to locate nucleation sites within an amino acid sequence. These nucleation sites are then extended until a stretch unlikely to belong to that structure is encountered, whereupon that structure is terminated. This process is repeated throughout the sequence until the secondary structure of the entire sequence is predicted.

The PHD method (Rost and Sander, 1992, Nature 360: 540) utilizes a combination of evolutionary and multiple sequence alignment information, and a "jury" of 12 networks. Since this method is a fully automated computer program, it is independent of human input or interpretation and as such delivers a unique approach.

2.2. Structure Of Glucose Transport Proteins

Mammalian glucose transporter proteins (GLUTS) constitute a family of proteins which are integrally embedded in the cell membrane and primarily transport glucose into and out of cells. Recent evidence indicates that compounds other than glucose, for example, water, dehydroascorbic acid and nicotinamide, can traverse GLUTs suggesting that these proteins may be multifunctional.

For example, glucose transporter proteins have recently been shown to exhibit a modest permeability to water (Fischbarg et al., 1990, Proc. Natl. Acad. Sci. USA, 87: 3244-3247), suggesting that the r is a channel in glucose transporter proteins that is hydrated and may serve as a conduit for the substrates mentioned in the paragraph above. Furth r, GLUT proteins may play a

role in the pathogenesis of diab tes, in that insulin elicits a specific and rapid response from GLUT proteins in human muscle and fat cells where a rapid translocation of GLUT from an internal storage pool to the plasma membrane occurs, thereby increasing the glucose uptake by these cells. In adipocytes, the Km for glucose may also be lowered as a response to insulin.

The GLUT proteins have been well characterized biochemically and their primary structures have been determined. But as is the case with many membrane proteins, the secondary structures of GLUTs are largely unknown, greatly hindering any study of their molecular mechanisms.

The hitherto most favored model of GLUT secondary structure predicts that GLUT proteins form 12 transmembrane α-helices (12H model; Mueckler et al., 1985, Science, 229:941-945). Further studies suggesting a high α-helical content include Chin et al., 1986, J. Biol. Chem. 261: 7101-7104 (Fourier transform infrared spectroscopy, FTIR) and Chin et al., 1987, Proc. Natl. Acad. Sci. U.S.A. 84: 4113-4116 (circular dichroiism, CD). Other studies have suggested that extensive α-helical content is accompanied by significant β-folding (FTIR spectroscopy: Alvarez et al., 1987, J. Biol. Chem. 262: 2502-3509; CD: Park et al., 1992, Protein Science 1;1032-1049), but have failed to appreciate the full extent of the β-structure predicted by the present invention.

The 12H model indicates that the highly conserved sequence (Ile 386 - Ala 405), in a particular GLUT protein, GLUT1, is intrac llular. However, recent experiments (Fischbarg et al., 1993, Proc. Natl. Acad. Sci. U.S.A. 90: 11658-11662) utilizing a synthetic polyclonal antibody to this conserved regi n showed that the antibody induced an increased glucose uptake

only when administ red extracellularly. This is inconsistent with the purported intracellular location of the region in the 12H model. These data, contradicting the established model, prompted further analysis of GLUT secondary structure using the novel algorithm of the invention and, as set forth below, led to the discovery of a new model for GLUT structure.

3. Summary Of The Invention

The present invention relates to methods of predicting the tendency of a portion of a protein to form amphiphilic α or β structure. It is based, at least in part, on the discovery that porin membrane proteins, which were previously assumed to contain predominantly α amphiphilic structure, unexpectedly are predicted to contain substantial amounts of β structure.

The methods of the present invention provide a number of advantages relative to methods previously used to analyze protein structure. For example, the CFP algorithm fails to consider hydrophobicity and amphiphilicity, and is more ambiguous in its predictions than the algorithms of the present invention. The CFP peaks are not fully representative of the actual protein structure, whereas the peaks seen by the Union program may provide a better visual representation of actual secondary structure.

In particular embodiments, the methods of the invention may be used to predict the presence of β-barrel structures in membrane proteins. The prediction of such structures in the protein may then be used for the rational design or identification of compounds that may interact with th pr tein. Alternativ ly, the methods of the invention may b used to create β-barrel structures in genetically ngin ered prot ins.

4. Description Of Th Figures

FIGURE 1 (D.E.H. and I) Data represent averaged values from two 10-oocyte groups: other data are averages from three such groups. Individual values differed with each other by <20%. For 60 min before the uptake assay, one group of oocyte (intracellular Ab, solid bars) was injected with 20-30 nl of a solution containing either Ab-1, Ab-4, or Ab-c (1 ng of Ab per 1 nl of water). A second group of oocytes (extracellular Ab, shaded bars) was incubated for 60 min in MBS containing the same ABs before measuring ^{3}H-DOG uptake. Control oocytes (open bars) were incubated in MBS. (D) Oocytes were incubated for 60 min with Ab in the outside incubation medium; the Ab concentration was varied as indicated. Solid circles, Ab-c; open circles (controls) Ab-4. (E) Oocytes incubated with Ab-c plus the addition of various concentrations of a peptide. The following peptides were used: solid circles, the conserved peptide Ile-386-Ala-405; open circles, the last 20 amino acids at the C-terminal end of GLUT4 (F) Oocytes incubated with Abs in the outside medium. Solid circles, Ab-c; open circles, Ab-4 (G) Open circles, oocytes incubated initially in medicum containing 1 μM insulin; arrow, the medium was replaced by another one containing insulin plus Ab-c (100 μg/ml). Solid circles, Ab-c in the initial incubation medium. Ab-c plus insulin after the arrow (H and I) Lineweaver-Burk plots of ^{3}H-DOG uptake in oocytes expressing GLUT1 and GLUT4, respectively, and incubated in the following media: open circles, MBS (controls); solid circles, MBS plus Ab-c (100 μg ml); triangles, MBS plus 1 μM insulin.

FIGURE 2. Multipl sequence alignment of two porins (OmpF, SEQ ID NO:1 and S16070, SEQ ID NO:2) and GLUT1 (SEQ ID NO:3). S16070 stands f r POR. Rectangles, existing (OmpF, POR) and predicted (GLUT1)

β-strands. Rounded rectangles, existing (OmpF, POR) and predicted (GLUT1) α-helices.

FIGURE 3. Prediction of porin structures using Union. Area graphs, $U_{\beta 7}$ prediction profiles. Structures known from crystallography (cryst.) or predicted (prd) are shown above the profiles in each case.

FIGURE 4. Our prediction for GLUT1. From top down, prediction profiles of hydrophobicity; turn; and union propensity for amphipathic α-helices and β-strands, respectively. Spans: 4 for <pt>: others are indicated in label subindices. For comparison, predicted structures are shown at the top and bottom panels. For the 12H model and for our prediction symbols are shown angled so that their lower and higher ends correspond to their intra-and extracellular sides, respectively.

FIGURE 5. Putative βB of GLUT1 viewed from inside the cell. A molecule of β-D-glucopyranose is shown in the center of the pore as a size marker (viewed from C1)

FIGURE 6. Model of secondary structure of GLUT1 (SEQ ID NO:3). Putative 16 transmembrane β-strands are represented by rectangles. The more hydrophilic sides of the β-strands (presumably lining the pore) are facing right. In the extramembrane loops, triangles denote predicted turns, and rectangles mark predicted α-helices. Of the two possible N-linked glycosylation sites N 45 and N 411, mutagenesis points to the first (Asano et al., 1991, J. Biol. Chem. 266:24632-24636). Two epitopes, 217-272 and 386-405 and a sugar binding site, Q 282 (Hashiramoto et al., 1992, J. Biol. Chem., 267: 17502-17507), are boldface.

FIGURE 7. Union profiles ar shaded. Cry: information from high-resolution structures. t: turns. (A) reaction center, L chain; (B) bacteriorhodopsin;

(C) colicin A; (D) *Rhodobacter capsulatus* porin; and (E) *Escherichia coli* porin."

FIGURE 8. Arrows mark predicted β-strands. (A) facilitative glucose transporter 1; (B) CHIP28; (C) acetylcholine receptor α-subunit; (D) lactose permease; (E) Na^+/glucose cotransporter; (F) shaker K channel; (G) calcium ATPase (sarcoplasmic reticulum); and (H) H^+/K^+ ATPase. In 8a, 4th panel, the dotted lines suggest the topological

orientation of predicted β-strands (intracellular at bottom). In 8d, panel 1, the alkaline phosphatase activity reported for the different fusions is superimposed on the H_{21} plot. Number labels identify the fusions. Also shown there is the 12H model for lac permease. As in 8a, a topological orientation is suggested (intracellular at bottom).

FIGURE 9. Zscores have been normalized (scoreN) for sequence length as in Park et al., 1992, Protein Sci. 1:1032-1049, namely: scoreN=scoreorig/C*[1-exp(A*sequence length +B)]. For each of the two environments depicted in panels (a) and (b), the same set of 400 randomly chosen globular proteins was run to generate a baseline distribution of raw scores vs. sequence length.

5. Detailed Description Of The Invention

For clarity of presentation, and not by way of limitation, the detailed description of the invention is divided into the following subsections:

(i) proteins to which the inventive structural determination method may be applied;

(ii) the Union algorithm;

(iii) the UNION program; and

(iv) the utility of the invention.

5.1. Proteins To Which The Inventive Structural Determination Method May Be Applied

The methods of the present invention may be applied to any protein, in order to determine the propensity of portions of the protein to form α and β structures.

In preferred embodiments of the invention, the methods are applied to membrane proteins, particularly proteins involved in transporting compounds between the intracellular and the extracellular compartments. For

example, and not by way of limitation, th methods of the present invention may be applied to the following proteins and to each member of their respective families: GLUT proteins (including but not limited to erythrocyte glycophorin), bacterial porins (including OmpC, OmpF, NmpA, NmpB, NmpC and LamB, etc.), aquaporins, bacteriorhodopsin and the bacteriorhodopsin precursor, the reaction center L chain, colicin A , *Rhodobacter capsulatus* porin , and *E. coli* porin, the acetylcholine receptor α subunit, *lac* permease, sodium-glucose co-transporter, shaker potassium ion channel, sarcoplasmic reticulum calcium-ATPase, components of the sodium ion/ potassium ion pump, gap junction proteins, cytokine receptors, the multidrug resistance transporter, the cystic fibrosis conductance regulator and "band III" protein of the erythrocyte membrane.

5.2. The Union Algorithm

The present invention provides for a Union algorithm which is able to predict the presence of amphiphilic α and/or β structures in proteins, preferably membrane proteins, as set forth below.

The present invention provides for a method of predicting the tendency of a portion of a protein to form an amphiphilic α structure, said portion having a span of x residues, wherein x is any integer, comprising calculating a value for $U_{\alpha x}$ using the equation $U_{\alpha x} = H_x + \mu_{\alpha x} - <pt>$. H_x is the average hydrophobicity for a span of x residues using the Kyte-Doolitte scale. $\mu_{\alpha x}$ is the hydrophobic moment (span x) as calculated by the method set forth in Eisenberg et al., 1984, Proc. Natl. Acad. Sci. U.S.A. 81: 140-144, for α structures, the angle between on residue and the successive residue b ing that ass ciat d with α h lices, such as about 90-110°, and preferably 100°. <pt> is the positi n dependent turn propensity, as

calculated according to the m thod set forth in Prevelige and Fasman, 1989, in "Prediction of Protein Structure and the Principles of Protein Conformation", Fasman, ed., Plenum Press, New York, pp. 391-416 (assigned to residue 2 in a 4-point turn). For example, a value of <pt> of a tetrapeptide is calculated as pt= fi x fi+1 x fi+2 x fi+3 when "i" is the residue and f= bend frequencies in the four positions of the α-turn.

The present invention also provides for a method of predicting the tendency of a portion of a protein to form an amphiphilic β structure, said portion having a span of *x* residues, wherein *x* is any integer, comprising calculating a value for $U_{\beta x}$ using the equation $U_{\beta x} = H_x + \mu_{\beta x}$ - <pt>. H_x is the average hydrophobicity for a span of *x* residues using the Kyte-Doolitte scale. $\mu_{\beta x}$ is the hydrophobic moment (span *x*) as calculated by the method set forth in Eisenberg et al., 1984, Proc. Natl. Acad. Sci. U.S.A. 81: 140-144, for β structures, the angle between one residue and the successive residue being that associated with β-structures, such as about 150-210°, preferably 160°. <pt> is the position dependent turn propensity, as calculated by the method set forth in Prevelige and Fasman, 1989, in "Prediction of Protein Structure and the Principles of Protein Conformation", Fasman, ed., Plenum Press, New York, pp. 391-416) (assigned to residue 2 in a 4-point turn). For example, a value of <pt> of a tetrapeptide is calculated as pt= fi x fi+1 x fi+2 x fi+3 when "i" is the residue and f= bend frequencies in the four positions of the β-turn.

Odd number residues are generally chosen in assigning hydropathy values so that a given sum could be plotted above th mid-residue of the segment. In preferred, nonlimiting embodiments, the value of *x* is seven or twenty-one. These valu s are preferable

because a length of seven residues represents the shortest span that can be reliably used with a minimum of localized "noise". A larger span of twenty-one residues was also chosen since this represents the average length of membrane spanning α-helices.

Accordingly, the extent of α or β structure may be determined using the Union algorithm by calculating the values for U as set forth above, for a series of portions spanning the protein or a relevant part of its structure, graphically depicting the results of these calculations, and performing the following analyses:

The α or β structure of the segments are interpreted on the basis of height and width of peaks in the Uαx and Uβx profiles and the predicted hydrophobicity of the segments. The height is determined relative to a threshold. The threshold may be arbitrarily set at 0, or may be assigned a different value, depending on the protein being analyzed. For example, as in the case of the porins, (see Figure 3) peaks were considered to originate at a threshold of zero, and were taken to predict β-structure if their upper segment exceeded a threshold set at about 2 (1.83 in one case, 2.15 in another) in an scale set to range from -4.5 to +4.5 (see Example 6, below). The value of 2 was chosen because it best fit proteins, known to have β-structure, used for calibration.

Peaks wide enough to correspond to a segment of the amino acid sequence long enough to span the membrane as an α-helix (e.g. 18-22, preferably 20 or 21 residues) are predicted to be α structures. Peaks that are too narrow to correspond to a segment of the amino acid sequence long enough to span the membrane as an α-helix but which are wide enough to correspond to a segment of the amino acid sequence with the correct length to span the membrane as β-strands (e.g., as

short as 6 r sidues, pr ferably 9-14) are predicted to be β structures. Transmembrane segments of a protein may thereby be predicted to comprise either amphiphilic α -type or β -type, or both.

In preferred embodiments of the invention, the foregoing methods may be practiced employing the source code for the Union algorithm, as set forth in Section 5.2.1, below. The Uαx or the Uβx profiles generated using the source code in Section 5.2.1. give a graphic visualization of the Uαx or Uβx values, respectively, of the segments from one end of the protein to the other. This tracks the hydrophobic and hydrophilic regions relative to a universal midline.

For example, and not by way of limitation, the use of methods comprising the Union algorithm to identify β-barrel structure in various proteins is set forth in sections 5 and 7, below.

5.2.1. Source Code For The Union Algorithm

```
'/*  PROGRAM UNION
'/*  J. FISCHBARG, F. CZEGLEDY, P. ISEROVICH; OCT. 1992-oct. 1993 */
'/*
'/*  TO CALCULATE AVG. HYDROPHOBICITY, AMPHIPHILICITY AND TURN POTENTIAL
'/*   OUTPUT  COLUMNS FOR SYMPHONY OR ORIGIN
'/*  please do not use word "UNION" in program (PB3 has command UNION)
' TAKES INPUT FROM DEFAULT TEXT FILE, NAMELY:
'    DEFAULTS = path$ + "DEFAULT.TEX"
'   TAKES SEQUENCE DATA FROM:
fileinp$ = path$ + filein$ + ".seq"
' WRITES PUTPUT DATA TO:
fileout$ = path$ + fileou$ + ".dat"
' p1 = round(houtsperm(i), 2)     '
'' p2=round(amphioutalphan(i),2)
'' p3=round(amphioutbetan(i),2)
' p4 = round(ptseqn(i), 2)
' p5 = round(unnalphan(i), 2)
' p6 = round(unnbetan(i), 2)
' p7 = round(houtmhein(i), 2)
$STATIC
$string 1
DECLARE SUB NORMAL (DUM!(), nlen%, DUNN!())
DECLARE SUB UN (ptseqn!(), NOUTN!(), AMPHIOUTN!(), nlen%, UNN!())
common path$,filein$,fileinp$,nseq,houtsperm(1),ptseqn(1),unnalphan(1),_
unnbetan(1),houtmhein(1)
DEFINT I-N
naa = 20
NSEQ = 1500
nwin = 26
DIM symbols$(naa)
DIM seqh(NSEQ)
DIM seqn(NSEQ)
DIM seqt(NSEQ)
DIM seq$(NSEQ)
DIM vkyte(naa)
DIM pturn(naa)
DIM nwsize(naa)
DIM hout(NSEQ)
DIM houtmhein(NSEQ)
DIM amphiout(NSEQ)
'DIM houtmhein(nseq)              'normalized
DIM houtsperm(NSEQ)              'normalized
DIM AMPHIOUTBETAN(NSEQ)          'normalized
DIM AMPHIOUTALPHAN(NSEQ)         'normalize
DIM turnprop(NSEQ)
DIM turnpropn(NSEQ)              'normalized
DIM UNNALPHA(NSEQ)
DIM unnalphan(NSEQ)              'normalized
DIM UNNBETA(NSEQ)
DIM unnbetan(NSEQ)               'normalized
DIM turnpropa(NSEQ)
DIM DUM(NSEQ)
DIM DUMN(NSEQ)                   'normalized
DIM pt(20, 4)
DIM PTSEQ(NSEQ)
DIM ptseqn(NSEQ)                 'normalized
 aacodes$ = "ARNDCQEGHILKMFPSTWYV"
CLS
'PRINT " ENTER PATH  (WITH ...\)"
```

```
'INPUT PATHS
pathS = "C:\MM\BEAUFORT\"
DEFAULTS = pathS + "DEFAULT.TEX"
OPEN DEFAULTS FOR INPUT AS #5
INPUT #5, fileinS, fileouS, angalpha, angbeta, WHEL, span, UW
CLOSE #5
fileinpS = pathS + fileinS + ".seq"
fileoutS = pathS + fileouS + ".dat"
'Kyte-Doolittle scale
'      1,  2,  3,  4,  5,  6,  7,  8,  9,  10
'LIST: A,  R,  N,  D,  C,  Q,  E,  G,  H,  I
 DATA 1.8,-4.5,-3.5,-3.5,2.5,-3.5,-3.5,-0.4,-3.2,4.5
'           11,  12, 13, 14,  15,  16,  17,  18,  19, 20
'      L,   K,  M,  F,  P,  S,  T,  W,  Y,  V
 DATA  3.8,-3.9,1.9,2.8,-1.6,-0.8,-0.7,-0.9,-1.3,4.2
'     Chou-Fasman turn propensities - scaled to correspond to KD indices
'       1 ,  2 ,  3 ,  4 ,  5 ,  6 ,  7 ,  8 ,  9 ,  10
'       A ,  R ,  N ,  D ,  C ,  Q ,  E ,  G ,  H ,  I

DATA  -2.9 ,-0.5 , 4.5 , 3.7 , 1.4 ,-0.3 ,-2.3 , 4.5 ,-0.5 ,-4.5
'       11 ,  12 ,  13 ,  14 ,  15 ,  16 ,  17 ,  18 ,  19 ,  20
'       L ,  K ,  M ,  F ,  P ,  S ,  T ,  W ,  Y ,  V
DATA  -3.5 , 0.0 ,-3.4 ,-3.4 , 4.2 , 3.4 ,-0.5 ,-0.5 , 1.0 ,-4.3


start:
CLS
DUMMY = 0
PRINT "1. FILE NAME FOR INPUT "; fileinpS
PRINT "2. FILE NAME FOR OUTPUT "; fileoutS
PRINT "3. ANGLE FOR ALPHA/BETA STRUCTURES (HYDROPHOBIC MOMENT) "
PRINT "   ALPHA STRUCTURE= "; angalpha; "BETA STRUCTURE= "; angbeta
PRINT "4. WINDOW SIZE FOR MEMBRANE HELICES "; WHEL
PRINT "5. SPAN FOR UNION "; span
PRINT "6. WINDOW SIZE FOR SMOOTHING UNION "; UW
PRINT "9. TO END SESSION "
PRINT " TO CHANGE ANY SETTINGS ENTER CORRESPONDING NUMBER "
PRINT " IF YOU ARE READY TO CONTINUE PRESS ENTER "
 INPUT DUMMY
 SELECT CASE DUMMY
 CASE 1
   GOSUB FILENAMEINPUT
' CASE  2
'         GOSUB   FILENAMEOUTPUT
 CASE 3
  GOSUB MOMENTANGLE
 CASE 4
  GOSUB ALPHAWINDOW
 CASE 5
  GOSUB BETAWINDOW
 CASE 6
  GOSUB UNNWINDOW
 CASE 9
  GOTO salida
 CASE 0
  OPEN DEFAULTS FOR OUTPUT AS #6
  WRITE #6, fileinS, fileouS, angalpha, angbeta, WHEL, span, UW
  CLOSE #6
  GOSUB WORKING
 END SELECT
```

```
120      PRINT 3
121      GOTO start
122
123     FILENAMEINPUT:
124     SHELL "DIR " + path$ + "*.SEQ/W"
125     LOCATE 20, 1
126     PRINT "ENTER FNAME (ONLY) FOR INPUT SEQ; PROG ADDS DEF FILE
           TYPES .SEQ & .DAT"
127     INPUT filein$
128     fileou$ = filein$
129     fileinp$ = path$ + filein$ + ".seq"
130     fileout$ = path$ + fileou$ + ".dat"
131      RETURN
132
133
134     WORKING:
135     OPEN fileinp$ FOR INPUT AS #1
136     INPUT #1, sequence$
137     CLOSE #1
138     turninput$ = path$ + "inp.dat"
139      OPEN turninput$ FOR INPUT AS #3
140      FOR I = 1 TO 20
141        FOR m = 1 TO 4
142        INPUT #3, pt(I, m)
143        NEXT m
144      NEXT I
145      CLOSE #3
146     FOR n = 1 TO 20
147       READ vkyte(n)
148       symbols$(n) = MID$(aacodes$, n, 1)
149     NEXT n
150     FOR n = 1 TO 20
151       READ pturn(n)            'acquire Chou-Fasman turn potentials
152     NEXT n
153     RESTORE
154     PRINT "..............  WORKING................."
155      nlen = LEN(sequence$)
156     FOR n = 1 TO nlen
157     seq$(n) = MID$(sequence$, n, 1)
158     NEXT n
159       ' end of while loop
160            FOR I = 1 TO nlen      ' from 1 to length of sequence */
161       FOR k = 1 TO 20
162       IF seq$(I) = symbols$(k) THEN ' identify ordinal FOR aa */
163         seqh(I) = vkyte(k)       ' and assign hydrophobicity value to residue*/
164         seqt(I) = pturn(k)
165         seqn(I) = k        'assign  residue name number*/
166       END IF
167         NEXT k
168         NEXT I
169     'cls
170     '       PRINT "A.A.: " seq$(i), "Hydr. i.= " seqh(i)
171     'j=5
172     ' FOR i=1 TO (nlen-j+1)  'i: beginn. of window; calcs.
              'all residues in wind.*/
173     ''       turnacc = 0
174      'FOR i=i TO (i+j-1)              'loop on i through all res. in window  /
175     '                turnacc = turnacc   seqt(i)
176      'NEXT i
177
178     '       m = i-1 (j+1)/2          ' define center of window */
179     '       turnprop(m) = turnacc/j 'calculate average turn propensity
```

```
180     ' NEXT l
181     FOR n = 2 TO (nlen - 2)
182     PTSEQ(n) = pt(seqn(n - 1), 1) * pt(seqn(n), 2) * pt(seqn(n + 1), 3)
         * pt(seqn(n + 2), 4)

183     NEXT n
184     PTSEQ(1) = PTSEQ(2)
185     PTSEQ(nlen) = PTSEQ(nlen - 2)
186     PTSEQ(nlen - 1) = PTSEQ(nlen - 2)
187     CALL NORMAL(PTSEQ(), nlen, ptseqn())
188
189       ' HYDROPHOBICITY CALCULATION FOR MEMBRANE HELICES * /
190     FLAG = 1 ' calculate hydrofobicity* /
191     j = WHEL                ' window
192     GOSUB MAIN      ' and we will get hout(m)
193     CALL NORMAL(hout(), nlen, houtmhein()) 'and we will get houtmhein*/
194
195       ' HYDROPHOBICITY CALCULATION FOR SHORT SPAN
196     FLAG = 1 ' calculate hydrofobicity* /
197     j = span                ' window
198     GOSUB MAIN      ' and we will get hout(m)
199     CALL NORMAL(hout(), nlen, houtspann()) 'and we will get houtspann*/
200
201       'calculation alpha moment*/
202     FLAG = 0
203     j = span
204     ANGLE = angalpha
205     GOSUB MAIN                      'and we will get amphiout(m)*/
206     CALL NORMAL(amphiout(), nlen, AMPHIOUTALPHAN())'and we will get
          AMPHIOUTALPHAN*/
207
208       ' calculation beta moment*/
209     FLAG = 0 'gives amphiout output
210     j = span
211     ANGLE = angbeta
212     GOSUB MAIN      'we get beta moment, amphiout(m)*/
213     CALL NORMAL(amphiout(), nlen, AMPHIOUTBETAN())  'an
214
215     'calculate union alpha
216     CALL UN(ptseqn(), houtspann(), AMPHIOUTALPHAN(), n
217     CALL NORMAL(UNNALPHA(), nlen, unnalphan())
218     'calculate union beta
219     CALL UN(ptseqn(), houtspann(), AMPHIOUTBETAN(), n
220     CALL NORMAL(UNNBETA(), nlen, unnbetan())
221
222     GOSUB producto
223     PRINT 5
224     RETURN
225     'GOTO  SALIDA
226
227     MOMENTANGLE:
228     PRINT "ENTER ANGLE FOR ALPHA STRUCTURES "
229     INPUT angalpha
230     PRINT "ENTER ANGLE FOR BETA STRUCTURES "
231     INPUT angbeta
232       RETURN
233
234     ALPHAWINDOW:
235     PRINT "ENTER WINDOW SIZE FOR  MEMBRANE HELICAL SPANS (ODD NUMBER)"
236     INPUT WHEL
237       RETURN
238
239     BETAWINDOW:
240     PRINT "ENTER WINDOW SIZE FOR  UNION SPAN (ODD NUMBER)"
```

```
241     INPUT span
242       RETURN
243
244     UNWINDOW:
245      PRINT "ENTER WINDOW SIZE FOR SMOOTHING UNION"
246     INPUT UW
247       RETURN
248
249     MAIN:       'window size j is already defined
250     IF j > 1 THEN
251     '*************** STARTING SEGMENT
252       FOR m = 1 TO (j - 1) / 2
253       LB = 1          'LOW BOUNDARY
254       UB = m + (j - 1) / 2'upper BOUNDARY
255       GOSUB CALCULATION
256       NEXT m
257     '******* END SEGMENT
258       FOR m = (nlen + 1 - (j-1)/2) TO nlen 'm  ctr of window*/
259       LB = m - (j - 1) / 2'low BOUNDARY
260       UB = nlen
261       GOSUB CALCULATION
262       NEXT m
263     ' MAIN CENTER SEGMENT
264       FOR m = (j + 1) / 2 TO (nlen - (j-1)/2)' m  center of the window*/
265       LB = m - (j - 1) / 2
266       UB = m + (j - 1) / 2
267       GOSUB CALCULATION
268       NEXT m
269     END IF
270       RETURN
271
272     CALCULATION:
273     IF FLAG = 1 THEN ' calculate hydrophobicity*/
274         cumh = 0: cum = 0                    ' reset hydrophobicity accumulator */
275       FOR I = LB TO UB                        'loop on i through all res. in window */
276         cumh = seqh(I) + cumh
277         cum = cum + 1
278         hout(m) = cumh / cum          ' compute hydrophobicity average *
279       NEXT I
280     ELSE                                      ' hydrophobic moment*/
281       t = 0
282      acum = 0
283      Mx = 0!
284      My = 0!          'reset amphi accumulators*/
285       FOR I = LB TO UB                                  'loop on i through all res.
            in window */
286      x = COS(2 * 3.1416 * ANGLE * (I - LB) / 360)'Eisinberg
287      y = SIN(2 * 3.1416 * ANGLE * (I - LB) / 360)
288      Mx = Mx + (x * seqh(I))
289      My = My + (y * seqh(I))
290
291      acum = acum + 1
292         NEXT I
293       amphiout(m) = SQR(Mx ^ 2 + My ^ 2)
294     END IF
295       RETURN
296     '       sub SMOOTH(dum(1),nlen,dums(1))
297     '       FOR    m=1 TO nlen
298     '       if  0<m< ((j-1)/2+1) then                'J window siz  for smoothing*
```

```
299     '       LB=1                'low boundary
300     '       UB=m+(j-1)/2        'upper boundary
301     '       end if
302     '                           NEXT m
303     '       FOR m=m+1-(j-1)/2  TO nlen ' m  center  f the window"/
304     '       LB=nlen-(j-1)/2    'low boundary
305     '       UB= nlen
306     '       GOSUB SMTH
307     '
308     '       FOR   m=(j+1)/2 TO (nlen-(j-1)/2) ' m  center of the window"/
309     '        LB=m-(j-1)/2
310     '        UB=m+(j-1)/2
311     '        GOSUB smth
312     '        NEXT m
313     '       return
314     SMTH:
315           SEQACCUM = 0: ACCUM = 0'RESET
316       FOR I = LB TO UB
317         SEQACCUM = SEQACCUM + DUM(I)
318         ACCUM = ACCUM + 1
319       NEXT I
320         DUMS(m) = SEQACCUM / ACCUM
321       RETURN
322     producto:
323     OPEN fileout$ FOR OUTPUT AS #2
324     'PRINT #2, "res", "H21", "H7", "m", "mb", "pt", "ub", "ub"
325     PRINT #2, "res", "H21", "H7", "pt", "ub", "ub"
326      FOR l = 1 TO nlen
327      PRINT "l= "; l; "; nlen= "; nlen
328      p1 = round(houtspenn(l), 2)
329     ' p2=round(amphioutalphan(l),2)
330     ' p3=round(amphioutbetan(l),2)
331      p4 = round(ptseqn(l), 2)
332      p5 = round(unnalphan(l), 2)
333      p6 = round(unnbetan(l), 2)
334      p7 = round(houtwheln(l), 2)
335     LOCATE 15, 1
336     'PRINT l;p7;p1;p2;p3;p4;p5;p6
337     PRINT l; p7; p1; p4; p5; p6
338     'print #2, "res#","N"UNEL,  "N"span, "m", "mb", "<pt>", "ub", "ub"
339     'PRINT #2,  l,      p7,      p1,      p2,   p3,    p4,    p5,   p6
340     PRINT #2, l, p7, p1, p4, p5, p6
341      NEXT l 'end do next row
342     CLOSE #2
343     RETURN
344     salida:
345     PRINT " ### DONE ####"
346
347           SUB NORMAL (DUM(), nlen, DUMN())
348
349       seqmax = DUM(1): seqmin = DUM(1): SEQCUM = DUM(1)' reset max and min and
        average accumulators"/
350       FOR I = 2 TO nlen'MAX AND MIN DETERMINATION  '"/
351         IF seqmax < DUM(I) THEN seqmax = DUM(I)
352         IF seqmin > DUM(I) THEN seqmin = DUM(I)
353         SEQCUM = SEQCUM + DUM(I)
354           NEXT I
355         SEQAVG = SEQCUM / nlen 'average
356           FOR I = 1 TO nlen
357         DUMN(I) = -4.5 + 9 * (DUM(I) - seqmin) / (seqmax - seqmin)
358             NEXT I
359       END SUB
```

```
360
361       SUB UN (ptseqn(), HOUTN(), AMPHIOUTN(), nlen, UNN())
362
363       FOR m = 1 TO nlen
364       UNN(m) = HOUTN(m) + AMPHIOUTN(m) - ptseqn(m)
365       NEXT m
366       END SUB
```

5.3. The Union Program

In preferred embodiments of the invention, the Union algorithm may be employed together with certain features of the Chou-Fasman-Prevelige, Kyte-Doolittle and PHD algorithms. This combination is referred to herein as the UNION program, source code for which is set forth in 5.3.1., below. For purposes of formatting, in a few instances, where the material for one line of source code could not fit into the margins, it was slightly indented and moved to the next line.

For example, and not by way of limitation, the following method may be performed.

(1) *Values of* H_x , $\mu_{\beta x}$ and <pt> and the obtained values of $U_{\beta x}$ may be calculated as set forth in the preceding section and their ranges scaled from -4.5 to +4.5 (see example sections 6 and 7).

(2) The Union algorithm may be used to mark the approximate location of the secondary structures. The Uαx or the Uβx profiles give a graphic visualization of the Uαx or Uβx values of the segments, respectively, relative to a universal midline.

(3) The α or β structure of the segments may be interpreted from the Uαx or Uβx profiles so that the width of the peaks from either profile may be compared to the actual distance needed to bridge a membrane. The segments, and thus the protein may be assigned an α or β structure based on the length of peaks in the Uαx and Uβx profiles and the predicted hydrophobicity of the segments.

(4) The segments may be refined using the CFP algorithm, as set forth in Prevelige and Fasman, 1989, in "Prediction of Protein Structure and th Principles

of Protein Conformation", Fasman, d., Plenum Press, New York, pp. 391-416, to calculate the values for α and β average propensities for tetrapeptides.

(5) Data from the neural network program PHD (Rost and Sander, 1992, Nature 360: 540) may be added as separate profiles of the segments.

(6) The various plots obtained from the methods described in (1) - (5) may be combined in a single figure for the global picture of an individual protein. This step renders the data maximally informative and is specified by the UNION program, the source code of which is set forth in Section 5.3.1, below. The UNION program runs in the IBM DOS or Microsoft-DOS environments, using a columnar input ASCII file that includes: (1) the amino-acid sequence of the protein and (2) a corresponding sequence of literal secondary structural assignment codes for that amino acid sequence, either from the Brookhaven database for proteins with known structure, or derived from predictions for proteins of unknown secondary structure. The literal structure codes are converted into numbers and a columnar output file is generated. Figures for data analysis may be conveiniently obtained by importing the UNION output into a graphics program: "ORIGIN", MicroCal Software, Northampton, MA 01060.

For example, and not by way of limitation, the use of the UNION program to identify β-barrel structure in various proteins is set forth in section 7, below.

5.3.1. Source Code For The Union Program

```
'/* PROGRAM UNION (for Union, and Chou-Fasman-Prevelige)
'/* J. FISCHBARG, F. CZEGLEDY, P. ISEROVICH, J. LI; COPYRIGHT 1994 */
'/* COLUMBIA UNIVERSITY, NEW YORK
'/* TO CALCULATE AVG. HYDROPHOBICITY, AMPHIPHILICITY AND TURN
      POTENTIAL
'/* OUTPUT COLUMNS FOR SYMPHONY OR ORIGIN
'/* please do not use word "UNION" in program (PB3 has command UNION)
'     TAKES INPUT FROM DEFAULT TEXT FILE, NAMELY:
'  DEFAULT$ = path$ + "union.INI"
'     TAKES SEQUENCE DATA FROM:
'     fileinp$ = path$ + filein$ + ".sqt"
' TAKES STRUCTURE INFORMATION FROM SAME SQT INPUT FILE:
' either from crystallog. or from preds., e.g., from PHD robot prediction
'     WRITES OUTPUT DATA TO:
'     fileout$ = path$ + "\"+ fileou$ + ".dat"
' Columnar output generated is:
'1) res: residue number
'2) aa: amino acid code
'3) H21: Kyte-Doolittle hydrophobicity, span selected for large windows
'(usually 21), assigned to center residue
'4) H7:  Kyte-Doolittle hydrophobicity, span selected for small windows
'(usually 7), assigned to center residue
'5) ua: Union for alpha structures, small window span (same as in H7)
'6) ub: Union for beta, etc.
'7) Pa: Chou-Fasman avg. alpha propensity for tetrapeptide (i,i+1,i+2,i+3)
'8) am: marker for supratheshold CF alpha (4.5 value for ease of plotting)
'9) Pb: CF avg.  beta propens. f tetrap.
'10) bm: marker for suprath. CF beta (4.5)
'11) pt: Chou-Fasman position-dependent tetrapeptide turn propensity
'(ass. to second residue)
'12) tm: marker for suprathreshold CF turn propensity (4.5)
'13) prda: alpha prediction marker (3.5 value)
'14) prdb: beta pred.  marker (3.5)
'15) prdt: turn pred. marker (3.5).

'The last three lines merely represent the conversion of the information in
'the second line of the input file.
' ////////////////////////////////////////////////////////////
$string 2
$static
cls

'DECLARE SUB NORMAL (DUM!(), nlen%, DUMN!())
'DECLARE SUB UN (ptseq!(), HOUT!(), AMPHIOUT!(), nlen%, UNN!())
'cls
```

```
common path$,filein$,fileinp$,nseq,houtsh(1),ptseq(1),ualpha(1),_
ubeta(1),hout(1)
DEFINT I-N
print "ENTER MAXIMUM SEQUENCE LENGTH TO DIMENSION ALL ARRAYS
     BY"
print "(preferably <2000; if more, might be limited by memory)"
PRINT " Default = carriage return = 1999"
input  nseq
if nseq=0 then nseq=1999
naa = 20    'unless dealing with extraterrestrials...
'NSEQ = 300  'maximum number of amino acids in sequence; sets array sizes
'       program sensitive to this in the Power Basic environment
'  however, if compiled, so far no limit encountered for the executable
DIM symbols$(naa)
DIM seqh(NSEQ)
DIM seqn(NSEQ)
DIM seq$(NSEQ)
DIM vkyte(naa)
DIM pturn(naa)
DIM hout(NSEQ)
DIM houtsh(NSEQ)
DIM amphi(nseq)
DIM amphiout(NSEQ)
DIM amphibeta(NSEQ)
DIM amphialpha(NSEQ)
DIM UALPHA(NSEQ)
DIM U(nseq)
DIM ubeta(NSEQ)
DIM DUM(NSEQ)
DIM DUMN(NSEQ)
DIM pai(naa)        'alpha propens. for individual amino acids
DIM pas(nseq)       'sequential indiv. alpha propens. along chain
DIM patetr(nseq)
DIM pam$(nseq)
DIM pbi(naa)        'beta propens. for individual amino acids
DIM pbs(nseq)       'sequential indiv. beta propens. along chain
DIM pbtetr(nseq)
DIM pbm$(nseq)
DIM pt(naa, 4)
DIM PTSEQ(NSEQ)
DIM ptm$(nseq)
DIM phda$(nseq)
DIM phdb$(nseq)
DIM phdt$(nseq)
DIM temp$(nseq)
DIM apos$(nseq)
```

```
DIM aneg$(nseq)
DIM aro$(nseq)

comienzo:
aacodes$ = "ARNDCQEGHILKMFPSTWYV"
alphacut = 100
betacut = 100
turncut = 0.75e-4
CLS
drive$ = "c:"
path$ = "\union"
DEFAULT$ = drive$ + path$ + "\" + "union.INI"
OPEN DEFAULT$ FOR INPUT AS #5
INPUT #5, drive$,path$,filename$,filename$,angalpha,angbeta,WHEL,span
CLOSE #5
fileinp$ = drive$ + path$ + "\" + filename$ + ".sqt"
fileout$ = drive$ + path$ + "\" + filename$ + ".dat"
'*************************************************************************
'Kyte-Doolittle scale
'    1,  2,  3,  4,  5,  6,  7,  8,  9, 10
'    A,  R,  N,  D,  C,  Q,  E,  G,  H,  I
DATA 1.8,-4.5,-3.5,-3.5,2.5,-3.5,-3.5,-0.4,-3.2,4.5
'    11, 12, 13, 14, 15, 16, 17, 18, 19, 20
'    L,  K,  M,  F,  P,  S,  T,  W,  Y,  V
DATA 3.8,-3.9,1.9,2.8,-1.6,-0.8,-0.7,-0.9,-1.3,4.2
'*************************************************************************
'CHOU-FASMAN 64-protein database
'*************************************************************************
DATA A,139, 79,0.060,0.076,0.035,0.058
DATA R,100, 94,0.070,0.106,0.099,0.085
DATA N, 78, 66,0.161,0.083,0.191,0.091
DATA D,106, 66,0.147,0.110,0.179,0.081
DATA C, 95,107,0.149,0.053,0.117,0.128
DATA Q,112,100,0.074,0.098,0.037,0.098
DATA E,144, 51,0.056,0.060,0.077,0.064
DATA G, 64, 87,0.102,0.085,0.190,0.152
DATA H,112, 83,0.140,0.047,0.093,0.054
DATA I, 99,157,0.043,0.034,0.013,0.056
DATA L,130,117,0.061,0.025,0.036,0.070
DATA K,121, 73,0.055,0.115,0.072,0.095
DATA M,132,101,0.068,0.082,0.014,0.055
DATA F,111,123,0.059,0.041,0.065,0.065
DATA P, 55, 62,0.102,0.301,0.034,0.068
DATA S, 72, 94,0.120,0.139,0.125,0.106
DATA T, 78,133,0.086,0.108,0.065,0.079
```

```
DATA W,103,124,0.077,0.013,0.064,0.167
DATA Y, 73,131,0.082,0.065,0.114,0.125
DATA V, 97,164,0.062,0.048,0.028,0.053

'       ////////////////////////////////////////////////////
CLS
'print "Free memory: ";fre(0); fre(-1); fre(-2)

PRINT "UNION ALGORITHM; J. Fischbarg, F. Czegledy, P. Iserovich, J. Li.
print" Set for sequence lengths up to  " nseq
print "
START:
DUMMY = 0
PRINT " ENTER ONE OF THE FOLLOWING "
PRINT
PRINT "1. CHANGE FILE NAME FOR INPUT; currently: "; fileinp$
'PRINT "2. CHANGE FILE NAME FOR OUTPUT; currently: "; fileout$
PRINT "2. CHANGE ANGLE FOR ALPHA/BETA STRUCTURES
      (HYDROPHOBIC MOMENT) "
PRINT "   ALPHA STRUCTURE= "; angalpha; "BETA STRUCTURE= "; angbeta
PRINT "3. CHANGE A.A. SPAN FOR MEMBRANE HELICES; currently: "; WHEL
PRINT "4. CHANGE A.A. SPAN FOR UNION; currently: "; span
PRINT "5. CHANGE PATH; CURRENTLY: " path$
print "6. CHANGE DRIVE; CURRENTLY: " drive$
'PRINT "6. CHANGE WINDOW SIZE FOR SMOOTHING UNION; currently: "; UW
PRINT "9. TO END SESSION WITHOUT RUNNING"
PRINT "0. (DEF=CR) MAIN - RUN WITH CURRENT PARAMETERS- RUNS ONLY
    ONCE AND EXITS
print
PRINT " "
 INPUT DUMMY
 SELECT CASE DUMMY
 CASE 1 :   GOSUB FILENAMEINPUT
'CASE 2 : GOSUB FILENAMEOUTPUT
 CASE 2 :  GOSUB MOMENTANGLE
 CASE 3 :  GOSUB ALPHAWINDOW
 CASE 4 :  GOSUB BETAWINDOW
 CASE 5 : GOSUB NEWPATH
 'TO COMIENZO   'road under repairs- monkeying with discouraged
'CASE 6 :  GOSUB UNNWINDOW
 CASE 9 :  GOTO salida
 CASE 0 : GOTO correte
END SELECT
GOTO start

'  ////////////////////////////////////////////////////
```

```
    producto:
    c$ = " , "
    OPEN fileout$ FOR OUTPUT AS #2
    PRINT #2, "res" c$ "aa" c$ "H21" c$ "H7" c$ "ua" c$ "ub" c$ "Pa" c$_
    "am" c$ "Pb" c$ "bm" c$ "pt" c$ "tm" c$ "prda" c$ "prdb" c$ "prdt" c$_
    "pos" c$ "neg" c$ "aro" c$
    FOR l = 1 TO nlen
    locate 16,1
    PRINT "l= "; l; "; nlen= "; nlen
          hlng = round(hout(l), 2)
          hsh = round(houtsh(l), 2)
          ua = round(ualpha(l), 2)
          ub = round(ubeta(l), 2)
          pa = round(patetr(l), 2)
          pb = round(pbtetr(l), 2)
          pt = round(ptseq(l), 2)
   PRINT #2,l c$ seq$(l) c$ hlng c$ hsh c$ ua c$ ub c$ pa c$ pam$(l)_
    c$ pb c$ pbm$(l) c$ pt c$ ptm$(l) c$ phda$(l) c$ phdb$(l) c$ phdt$(l)_
    c$ apos$(l) c$ aneg$(l) c$ aro$(l)
   NEXT l
   CLOSE #2                                                    'and do next row
   erase hout,houtsh,ualpha,ubeta,patetr,pbtetr,ptseq
   return

   '////////////////////////////////////////////////////////////
   newpath:
   print "Enter new path, e.g., C:\PROT\PRD(note no end \ or end space)"
   input test$
   if test$="" then goto newpath
   path$ = test$
   DEFAULT$ = drive$ + path$ + "\" + "union.INI"
   fileinp$ = drive$ + path$ + "\" + filename$ + ".sqt"
   fileout$ = drive$ + path$ + "\" + filename$ + ".dat"
   return
  '////////////////////////////////////////////////////////////
  correte:       'main routine - records parameters and runs
   cls
   OPEN DEFAULT$ FOR OUTPUT AS #6
   WRITE #6,drive$,path$,filename$,filename$,angalpha,angbeta,WHEL,span
   CLOSE #6
   GOSUB WORKING
PRINT
print "COMPLETED RUN SUCCESSFULLY - STOPPING NOW"
STOP  'With all the erase statements to save memory, cannot run again
      'all key arrays erased by now
```

```
'GOTO start
'////////////////////////////////////////////////////////////////

FILENAMEINPUT:
cls
chdrive drive$
chdir path$
files  "*.SQT"
files  "*.DAT"
LOCATE 20, 1
PRINT "ENTER FNAME (ONLY) FOR INPUT SEQ; PROG ADDS DEF FILE
      TYPES .SQT & .DAT"
INPUT filename$
fileinp$ = drive$ + path$ + "\" + filename$ + ".sqt"
fileout$ = drive$ + path$ + "\" + filename$ + ".dat"
 RETURN

MOMENTANGLE:
PRINT "ENTER ANGLE FOR ALPHA STRUCTURES "
INPUT angalpha
PRINT "ENTER ANGLE FOR BETA STRUCTURES "
INPUT angbeta
 RETURN

ALPHAWINDOW:
PRINT "ENTER WINDOW SIZE FOR  MEMBRANE HELICAL SPANS (ODD
      NUMBER)"
INPUT WHEL
 RETURN

BETAWINDOW:
PRINT "ENTER WINDOW SIZE FOR  UNION SPAN (ODD NUMBER)"
INPUT span
 RETURN

UNNWINDOW:
 PRINT "ENTER WINDOW SIZE FOR SMOOTHING UNION"
INPUT UW
 RETURN

' ////////////////////////////////////////////////////////////////

WORKING:
print fre(0); fre(-1); fre(-2)
OPEN fileinp$ FOR INPUT AS #1
INPUT #1, sequence$
input #1, structure$
```

```
    CLOSE #1
    FOR n = 1 TO 20
          READ vkyte(n)
          symbols$(n) = MID$(aacodes$, n, 1)
    NEXT n
    FOR i = 1 TO 20
    READ symbols$(i),pai(i),pbi(i),pt(I,1),pt(i,2),pt(i,3),pt(i,4)
    NEXT i
    RESTORE
                        '///////  **************      ////////////////////
    PRINT "..... WORKING....."
    cfspan = 4                    'prepare for Chou-Fasman-Prevelige tetrapeptides
    nlen = LEN(sequence$)
    FOR n = 1 TO nlen
    seq$(n) = MID$(sequence$, n, 1)    'list of aa codes
    NEXT n
    FOR I = 1 TO nlen     ' from 1 to length of sequence */
          FOR k = 1 TO 20
          IF seq$(I) = symbols$(k) THEN ' identify ordinal FOR aa */
                seqh(I) = vkyte(k)  '  assign hydrophobicity value to residue*/
                seqn(I) = k              'assign  residue name number*/
        pas(i) = pai(k)          'assign alpha propensity
        pbs(i) = pbi(k) 'assign beta propensity
        exit for                 'done here; leave for/next loop
         END IF
               NEXT k
 NEXT I

 FOR n = 2 TO (nlen - 2)
 PTSEQ(n)=pt(seqn(n-1),1)*pt(seqn(n),2)*pt(seqn(n+1),3)*pt(seqn(n+2),4)
 NEXT n
 erase seqn
PTSEQ(1) = PTSEQ(2)
PTSEQ(nlen) = PTSEQ(nlen - 2)
PTSEQ(nlen - 1) = PTSEQ(nlen - 2)
for i=1 to nlen
if ptseq(i) >= turncut then
   for ind = 0 to 3
       ptm$(i + ind) = "4.5 " : next ind : goto cortada
end if
if ptseq(i) < turncut then
       if ptm$(i) =  "4.5 " then goto cortada
       else
             ptm$(i) = "    "
end if
```

```
cortada:
next i
CALL NORMAL(PTSEQ(), nlen, ptseq())

'       ********** ////////////////// **********

 ' HYDROPHOBICITY CALCULATION FOR MEMBRANE HELICES * /
FLAG = 1 ' calculate hydrophobicity* /
j = WHEL          ' window
GOSUB MAIN     ' and we will get hout(m)
CALL NORMAL(hout(), nlen, hout()) 'and we will get hout long*/

'       ********** ////////////////// **********'

 ' HYDROPHOBICITY CALCULATION FOR SHORT SPAN
FLAG = 2 ' calculate hydrophobicity* /
j = span          ' window
GOSUB MAIN     ' and we will get houtsh(m)
CALL NORMAL(houtsh(), nlen, houtsh()) 'and we will get hout short*/
'       ********** ////////////////// **********
'       CALCULATION OF TETRAPEPTIDE PROPENSITIES
j = cfspan
for i=1 to nlen-3
patetr(i) = ( pas(i) + pas(i+1) + pas(i+2) + pas(i+3) )/cfspan
pbtetr(i) = ( pbs(i) + pbs(i+1) + pbs(i+2) + pbs(i+3) )/cfspan
if patetr(i) >= alphacut then
       pam$(i) = "4.5 "
else
       pam$(i) = "    "
end if
if pbtetr(i) >= betacut then
       pbm$(i) = "4.5 "
else
       pbm$(i) = "    "
end if
next i
 erase pas, pbs
for j= 2 to 0 step -1                    'approximate bottom ends
patetr(nlen-j) = patetr(nlen-3)
pbtetr(nlen-j) = pbtetr(nlen-3)
next j

CALL NORMALPA(patetr(), nlen, patetr() )
CALL NORMALPB(pbtetr(), nlen, pbtetr() )
```

```
'    calculation alpha moment*/

FLAG = 0  'selects amphiout output
j = span
ANGLE = angalpha
GOSUB MAIN          'gets amphiout(m)*/
CALL NORMAL(amphiout(), nlen, amphialpha())'gets amphialpha*/

'       ************ ////////////////// ***************
' calculation beta moment*/

FLAG = 0 'selects amphiout output
j = span
ANGLE = angbeta
GOSUB MAIN      'gets amphiout(m)*/
CALL NORMAL(amphiout(), nlen, amphibeta())  'gets amphibeta
 erase amphiout
'       ************ ////////////////// ***************

'calculate union alpha
CALL UN(ptseq(), houtsh(), amphialpha(), nlen, ualpha())
 erase amphialpha
CALL NORMAL(ualpha(), nlen, ualpha())
'calculate union beta
CALL UN(ptseq(), houtsh(), amphibeta(), nlen, ubeta())
 erase amphibeta
CALL NORMAL(ubeta(), nlen, ubeta())
 erase amphi
 erase seqh
'       ************ ////////////////// ***************
'       ////////////////////////////////////////////////////////////
'       PROCESS STRUCTURE STRING (PREDICTIONS OR CRYSTALLOG.)
alfam$ = "3.5 " : betam$ = "3.5 " : turnm$ = "3.5 "
posm$  = "2.5 " : negm$  = "2.0 " : arom$  = "5.5 "
FOR n = 1 TO nlen
temp$(n) = MID$(structure$, n, 1)  'list of structure codes
NEXT n
for i=1 to nlen
if temp$(i) = "H" then
      phda$(i) = alfam$  :  phdb$(i) = "    " : phdt$(i) = "    "
end if
if temp$(i) = "E" then
      phda$(i) = "    "  :  phdb$(i) = betam$ : phdt$(i) = "    "
end if
if temp$(i) = "C" then
```

```
        phda$(i) = "   " : phdb$(i) = "   " : phdt$(i) = "   "
 end if
 if temp$(i) = "T" then
        phda$(i) = "   " : phdb$(i) = "   " : phdt$(i) = turnm$
 end if
 if seq$(i) = "F" or seq$(i)="Y" or seq$(i)="W" then
        aro$(i) = arom$
   else
   aro$(i) = "   "
 end if
 if seq$(i) = "E" or seq$(i)= "D" then
        aneg$(i) = negm$
   else
   aneg$(i) = "   "
 end if
 if seq$(i) = "R" or seq$(i)= "K" then
        apos$(i) = posm$
   else
   apos$(i) = "   "
 end if

 next i
 close #4
 erase temp$
 GOsub producto
 return

' ////////////////////////////////////////////////////////////

 MAIN:     'window size j is already defined
 IF j > 1 THEN
 '*************** STARTING SEGMENT
       FOR m = 1 TO (j - 1) / 2     '1 to 10
       LB = 1        'LOW BOUNDARY
       UB = m + (j - 1) / 2'upper BOUNDARY  m+10
       GOSUB CALCULATION
       NEXT m
 ' ********MAIN CENTER SEGMENT   '11 to nlen - 10
       FOR m = (j + 1) / 2 TO (nlen - (j - 1) / 2)' m  center of the w
       LB = m - (j - 1) / 2  ' m-10
       UB = m + (j - 1) / 2  ' m+10
       GOSUB CALCULATION
       NEXT m
 '****  ** END SEGMENT          'nlen-9 to nlen
       FOR m = (1 + nlen - (j - 1) / 2) TO nlen 'm  ctr of window /
```

```
        LB = m - (j - 1) / 2 'low BOUNDARY  m-10
        UB = nlen
        GOSUB CALCULATION
        NEXT m
 END IF
  RETURN

' ////////////////////////////////////////////////////////////

 CALCULATION:

 IF FLAG = 1 THEN          ' calculate hydrophobicity of std. tm. segmts.*/
 cumh = 0: cum = 0         ' reset hydrophobicity accumulators */

 FOR I = LB TO UB          'loop on i through all res. in window */
 cumh = seqh(I) + cumh
 cum = cum + 1
 hout(m) = cumh / cum              ' compute hydrophobicity average *
 NEXT I

 ELSEIF FLAG = 2 THEN  ' calculate hydrophobicity of short tm. segmts.*/
 cumh = 0: cum = 0         ' reset hydrophobicity accumulators */

 FOR I = LB TO UB          'loop on i through all res. in window */
 cumh = seqh(I) + cumh
 cum = cum + 1
 houtsh(m) = cumh / cum            ' compute hydrophobicity average *
 NEXT I

 ELSEIF FLAG = 0 THEN          ' calc. hydrophobic moment*/

 t = 0 : acum = 0 : Mx = 0! : My = 0!    'reset amphi accumulators*/
 FOR I = LB TO UB        'loop on i through all res. in window */
 x = COS(2 * 3.1416 * ANGLE * (I - LB) / 360)   'Eisenberg
 y = SIN(2 * 3.1416 * ANGLE * (I - LB) / 360)
 Mx = Mx + (x * seqh(I))
 My = My + (y * seqh(I))
 acum = acum + 1
 NEXT I
 amphiout(m) = SQR(Mx ^ 2 + My ^ 2)

 END IF

  RETURN
```

```
'     ////////////// *     **  * *     *            */////////////////

salida:
PRINT " ### DONE ####"
stop
end


' ////////////////////////////////////////////////////////////////

    SUB NORMAL (DUM(), nlen, DUMN())

ytop#=DUM(1):ybot#=DUM(1):yCUM#=DUM(1)' reset max, min & avg accumulators*/
      FOR I = 2 TO nlen 'MAX AND MIN DETERMINATION '*/
            IF DUM(I) > ytop# THEN
          ytop# = DUM(I)
          yhord = i
          end if
            IF DUM(I) < ybot# THEN
          ybot# = DUM(I)
          ylord=i
          end if
'           yCUM# = yCUM# + DUM(I)
      NEXT I
'           yAVeraG = yCUM# / nlen 'average
      FOR I = 1 TO nlen
'print DUM(I)
'print (DUM(I) - ybot#);(ytop# - ybot#);(DUM(I) - ybot#)/ (ytop# - ybot#)
      DUMN(I) = -4.5 + 9 * (DUM(I) - ybot#) / (ytop# - ybot#)
      NEXT I

      END SUB

' ////////////////////////////////////////////////////////////////

  SUB UN (ptseq(), HOUTsh(), AMPHI(), nlen, U())

FOR m = 1 TO nlen
      U(m) = HOUTsh(m) + AMPHI(m) - ptseq(m)
NEXT m
 END SUB
      ************************************************
      SUB NORMALPA (DUM(), nlen, DUMN())
```

```
deltalfa# = 75
llalfa# = 64
  FOR I = 1 TO NLEN
     DUMN(I) = -4.5 + 9 * (DUM(I) - llalfa#) / (deltalfa#)
     NEXT I

     END SUB
     ************************************************************
       SUB NORMALPB (DUM(), nlen, DUMN())

deltabeta# = 106
llbeta# = 51
  FOR I = 1 TO NLEN
     DUMN(I) = -4.5 + 9 * (DUM(I) - llbeta#) / (deltabeta#)
     NEXT I

     END SUB
```

5.4. The Utility Of The Invention

The present invention provides for a method of predicting the structure of membrane proteins, which may be used in the following non-limiting embodiments.

In preferred embodiments, the method of the invention may be used to identify β-barrel structures in membrane proteins. The identification of β-barrel structure may be consistent with the function of the membrane protein as a translocator. As such, the present invention may be used to discern the function of membrane proteins, the function of which has been hitherto unknown.

Further, the identification of β-barrel structure in a protein may lead to the identification of molecules that can be transported by the protein. For example, the identification of a structure similar to members of the GLUT family of proteins in a particular protein would suggest that the protein may be able to translocate compounds similar to hexose compounds through a cell membrane containing that protein. Such an analysis may aid in the rational design of pharmaceutical agents that could be used to access a cell expressing the protein in its membrane.

In further embodiments, the present invention may be used to design or identify compounds able to be transported by animal or plant aquaporins (Chrispels and Agre, TIBS, 1994, : 421-425). In the case of animal aquaporins, the channel forming integral protein (CHIP), abundant in certain plasma membranes, and other homologs suggest that some of these proteins may be involved in clinical syndromes. Plant aquaporins like Tonoplast intrinsic protein (γ-TIP) can be used to study the role of these molecules in the water economy of plants, as well as to cr ate transgenic plants that express these proteins from tissue specific promoters. Drought-resistance and hardiness in crop plants may be

correlated with the presence and activity of these pr teins. The present invention can be used to address the current problems present in analyzing and manipulating the molecular structure and function of this family of membrane proteins.

In still further embodiments, the present invention may be used to engineer proteins having useful β-barrel structures. For example, the ability of a number of aquaporin proteins may be compared, and the particular protein having the most favorable transport capability may be identified. The method of the present invention may then be used to analyze its structure, and the secondary structures of other membrane proteins may be manipulated to resemble the structural characteristics of the designated aquaporin.

6. EXAMPLE: EVIDENCE THAT FACILITATIVE GLUCOSE TRANSPORTERS MAY FOLD AS B-BARRELS

6.1. MATERIALS AND METHODS

Antibody Studies. We raised three polyclonal antibodies ("Abs") in rabbits and used the IgG fractions. They were Ab-1, against the last 21 C-terminal amino acids of the GLUT1 protein; Ab-4 against the last 25 C-terminal amino acids of the GLUT4 protein (Ab-1 specifically reacted with GLUT1 but not with GLUT2 or GLUT4, and Ab-4 reacted with GLUT4 but not with GLUT1 or GLUT2 as assessed by immunoprecipitation and immunoblotting; and Ab-c raised against a synthetic peptide containing the sequence Ile-386-Ala-405 in GLUT1, a sequence that is highly conserved in all members of the GLUT family. Ab-c reacted with the GLUT1, GLUT2 and GLUT4 isoforms of mammalian facilitative transporters as assessed by immunoprecipitation and immunoblotting and the r activity was specifically blocked by competition with an exc ss of the peptide used to generate the Ab but

not by an unrelated peptid. For the experiments all Abs were suspended at a final concentration of 100 μg of IgG per ml in modified (Vera et al., 1990, Mol. Cell Biol., 10:743-751) Barth's solution (MBS).

Xenopus laevis oocytes were isolated as described (Vera et al., 1990, Mol. Cell Biol., 10:743-751) and injected with 50 nl of water containing 10-20 ng of *in vitro* synthesized capped RNA (Vera, supra.) encoding either GLUT1, GLUT2, or GLUT4, and incubated in MBS. Three days after RNA injection, uptake of 2-deoxy-[1.2(n)-^{3}H]D-glucose (^{3}H-DOG) was measured using a 10-min uptake assay (Vera, supra.). Oocytes were placed into 1 ml of MBS containing 0.5 mM DOG and 1-5 μCi of ^{3}H-DOG per ml (10 Ci/mmol: 1 Ci=37 GBq:NEN/DuPont). Ten pooled oocytes yielded an uptake value; values were consistent within a given batch of oocytes.

Alignments: We used the BESTFIT and PILEUP routines of the GCG (Genetics Computer Group; Version 7.0) program package, with gap weight = 3.0 and length weight = 0.1 (Needleman et al., 1970, J. Mol. Biol., 48:443-453). We aligned the sequences of *Rhodobacter capsulatus poria* (SEQ ID NO:2; Weiss et al., 1991, FEBS Lett., 280:379-382), *Escherichia coli* porin (SEQ ID NO:1; Sw; Ompf-Ecoli), and GLUT1 (SEQ ID NO:3; Sw:Gtr1-Human).

Predictions. We developed an algorithm ("Union") to predict protein segments with relatively high hydrophobicity and propensity to form amphiphilic α or β structures. For a residue span length i, Union (U) is: $U_{\iota i}=H_i-\mu_{\iota i}$-(pt) (Equation 1).

Depending on the structure for which U is calculated, the subindex ι stands for either α or β. H_i is the average hydrophobicity for a span of i residues using th Kyte-Doolittle scale (Kyte, et al., 1982, J. Mol. Biol., 157:105-132): $\mu\iota i$ is the hydrophobic moment (Eisenberg et al., 1984, Proc. Natl. Acad. Sci.

U.S.A., 81:140-144; span i) for either α or β structures: the angles between a residue and the next for α and β structures were 100° and 160°, respectively, using standard values for α-helices and the generic twist of β-sheets. H_i and $\mu_{i,i}$ values were assigned to the center residue of given odd-valued spans. <pt> is the position-dependent turn propensity (Prevelige, and Fasman, 1989, in "Prediction of Protein Sructure and the Principles of Protein Conformation", Fasman, ed., Plenum Press, New York, pp. 391-416; assigned to residue 2 in the 4-point turn). We calculated values of H_i, $\mu_{i,i}$, and <pt> for a given sequence and scaled their ranges to -4.5 to +4.5 in each case. After algebraic addition (Eq. 1), the $U_{i,i}$ values obtained were in turn rescaled to -4.5 to +4.5. We used union profiles to mark the approximate locations of secondary structures. Segments were then refined by using (i) the Chou-Fasman-Prevelige prediction method (CFP), which requires judgments by the operator, and (ii) the results from a neural network prediction program [PHD: profile neural network prediction, Heidelberg; Rost and Sander 1992, Nature, 360:540)], which runs unbiased, without human intervention. We found it convenient to display propensity profiles using the program PSAAM (Crofts, A.R., 1992, Ph.D. Dissertation (University of Illinois, Unknown). Three-dimensional modeling was done in the Insight and Discover graphical environments. (Biosym Technologies, San Diego).

6.2. Results And Discussion

Effect of Abs n the Function of Mammalian H x se Transport rs Expressed in *X. laevis* Oocyt s. Th highly conserved s quence (Ile-386-Ala-405 in GLUT1) is predict d to be intracellular in th 12H m del (Mueckler et al., 1985, Science, 229:941-945), which

locates it b tween its putative tm regions 10 and 11. Given the evidence for an important functional role for the region between tm domains 4 and 12 in GLUT1 (Carruthers, 1990, Physiol. Rev., 70:1135-1175), we reasoned that an Ab against that conserved sequence might elicit inhibition or activation of the transporter. After verifying its reactivity, we used *X. laevis* oocytes expressing different members of the mammalian GLUT family to study the effect of this anti-peptide Ab on the uptake of DOG. Incubation with Ab for 1 hour induced a measurable increase in the ability of oocytes expressing any of the three mammalian transporters tested, namely GLUT1 (Fig. 1A. c), GLUT2 (Fig. 1B. c), and GLUT 4 (Fig. 1C,c) to take up DOG. The Ab, however, acted only when present in the extracellular medium (Fig. 1A-C, c). No effect on uptake was observed when the Ab was injected into the oocytes 1 hr before the uptake measurements (Fig. 1A-C). The effect of Ab was dose dependent (Fig. 1D) and was specifically blocked by competition with excess peptide during the incubation period (Fig. 1E). The effect of the Ab on DOG uptake was evident after a short incubation period; near-maximal levels of activation were reached in ≈30 min (Fig. 1F). Incubation for several hours induced an additional increase in uptake (Fig. 1F).

To determine whether the GLUTs were expressed with the correct orientation in the membrane of the oocytes, we tested the effect of two other anti-peptide Abs we elicited against the C-terminal regions of GLUT1 and GLUT4. It was known from previous studies that this region of the transporters is located intracellularly (Oka et al., 1990, Natur , 345:550-553). As xpected, the Abs did not aff ct the capacity of th oocytes to take up DOG when added extracellularly (Fig. 1 A-C) but caus d a specific and m asurable increas in th

ability of oocytes expressing GLUT1 or GLUT4 (but not GLUT2), to take up DOG when injected intracellularly (Fig. 1 A-C). These observations are consistent with previous indicates that the C-terminal region is central to the function of the transporter (Oka, supra.).

Since both the Ab (Ab-c) and insulin (Vera, et al., 1990, Mol. Cell Biol., 10:743-751) increase DOG uptake in oocytes, we investigated whether Ab could act by mimicking insulin rather than by specifically binding to GLUTs. The results in Fig. 1 G-I suggest instead that the Ab and insulin have different mechanisms of action. Incubation of the oocytes with insulin did not affect the K_m of the transporters for DOG, increasing instead the V_{max} (Fig. 1H and I; Table 1). This is inconsistent with insulin inducing the translocation of transporters to the cell membrane. On the other hand, the Ab induced a measurable decrease in the K_m for DOG in oocytes expressing either GLUT1 or GLUT4 without changing the V_{max} (Fig. 1 H and I; Table 1). The short-term effect of the Ab on uptake (Fig. 1F) can be accounted for by an increased affinity of the transporters for DOG. The additional increase in uptake observed after long incubation periods with the Ab (Fig. 1F) may be due to the entrapment of the transporters at the level of the cell membrane.

Table 1 Effects of insulin and Ab on V_{max} and K_m values

Complementary RNA		V_{max}±SE. pmol per oocyte per min	K_m±SE mM
GLUT1 (from Fig. 1H)	Control	109±23	8.6±3.2
	Antibody	90±2	2.8±0.2
	Insulin	163±22	6.1±1.7
GLUT 4 (from Fig. 1M)	Control	76±10	8.0±1.9
	Antibody	60±2	2.7±0.2
	Insulin	119±9	7.8±1.1

Additional evidence for the different modes of action of the Ab and insulin came from experiments in which oocytes were first treated with insulin and then with the Ab and vice versa. Under the first condition, the Ab induced a further 2-fold increase in uptake in oocytes pretreated with insulin (for a total 4-fol increase; Fig. 1G). Quantitatively, this result is consistent with the effect of the Ab on the affinity of the transporter for DOG. On the other hand, insulin did not affect the uptake of DOG in oocytes previously treated with the Ab (Fig. 1G). One explanation for this finding is that the binding of the Ab to the transporter may "anchor" it to the plasma membrane and disrupt the dynamic equilibrium that allows insulin to modify the ratio of transporters located intracellularly versus those located at the plasma membrane.

The topology induced for the Ab findings compromises 12H. A possible explanation for the effect of Ab recognition of the sequence Phe-389-Ala-403 in terms of the 12H model is to argue that perhaps tm helices 10 and 11 are in a highly mobile segment of the protein, leading to the exposure of the internal loop between them to the extracellular medium. There is an α-helical membrane protein, colicin, which appears to externalize some of its α-helices during large scale conformational changes (Parker et al., 1992, J. Mol. Biol., 224:639-657). Externalization, however, shuts off the colicin channel, while in the present case uptake by GLUTs is enhanced by the Ab-c, militating against a colicin-type mechanism. Moreover, the Ab-c had no effect when injected intracellularly, further evidence against the intracellular location of Phe-389-Ala-403. The simplest explanation for our findings is that the loop comprising the segment Phe-389-Ala-403 is normally located on the extracellular side of the

membran , suggesting a t p l gy inconsist nt with the 12H model. If GLUTs are multihelical, with tm helices ≈20 residues long, and if putative helices 11 and 12 exist, then the converged loop could only be intracellular, being separated from the intracellular C-terminal loop by the hairpin of these two helices (see Fig. 4).

An alternative scheme: GLUT1 and the Porins. Given the foregoing, we searched for an alternative secondary structure for the transporter. We considered the structures of those few membrane proteins that have been solved by crystallography so far, and we came upon porins. In contrast to α-helical membrane proteins crystallized earlier, porin monomers form 16-stranded antiparallel βBs (Weiss et al., 1991, FEBS Lett., 280:379-382; Cowan et al., 1992, Nature, 358:727-733). When we aligned (Fig. 2) the sequences of *R. capsulatus* porin (POR; SEQ ID NO:2), *E. coli* porin (OmpF; SEQ ID NO:1), and GLUT1 (SEQ ID NO:3) (using BESTFIT), we found pairwise scores for identity and similarity as follows: POR-OmpF 20.0 and 45.7: POR-GLUT1, 19.9 and 46.6: OmpF-GLUT1, 18.2 AND 42.9. Porins in general show little overall primary sequence similarity (Welte et al., 1991, Biochim. Biophys. Acta, 1080:271-274). In particular, although the secondary structures of POR and OmpF are the same, the scores for the alignment are modest. The alignments of GLUT1 with the porins, however, elicit about the same scores as the alignment of the two porins. Hence, we set out to evaluate a possible porin-fold for GLUT1.

Prediction of Multiple tm β-Strands in Porins. From exploratory work, we chose a span of 7 residues to examine POR, OmpF and GLUT1 pr files. W found that the union β7 ($U_{\beta 7}$) peaks identified the approximate location and length of the β-strands in both porins (Fig. 3). The thresholds in Fig. 3 (1.83 for POR: 2.15

for OmpF) wer selected so as not to miss any strand; they result in only minimal overprediction. Segments were then refined by the CFP procedure. In comparing the porin structures thus predicted with those known from x-ray crystallography (Weiss et al., 1991, FEBS Lett., 280:379-382; Cowan et al., 1992, Nature, 358:727-733), we found success rates [Q3 (Qan, 1988, J. Mol. Biol., 202:865-884)] of 0.70 and 0.75 for POR and OmpF, respectively. The correlation coefficients (Mathews, 1975, Biochim. Biophys. Acta, 405:442-451) for our predictions were as follows -- for POR: α 0.56; β 0.70; turns. 0.28; random. 0.48; for OmpF; α 0.25; β 0.64; turns. 0.30; random 0.44. The PHD method (available only for OmpF) predicted regions with secondary structure similar to ours Q3 = 0.68).

Prediction of Multiple tm β-Strands in GLUT1. We identified 16 predicted tm β-strands in GLUT1 (Fig. 4). All were in segments that had been allocated as tm helices in the 12H model (Fig. 2). Using only H_{21} profiles, several of the peaks seen (Fig. 2) appeared wide enough to be interpretable as tm α-helices with spans of 21 residues (Mueckler et al., 1985, Science, 229:941-945). However, four of them (arrows in Fig. 4) were split by predicted turns. The resulting segments were too short to bridge the membrane as α-helices but had the correct length for tm β-strands. We termed such patterns "β-hairpin signatures." Similarly, in the remaining 8 segments previously predicted as 20-residue helices (Fig. 2) we predicted tm β-strands approximately 10 residues long, with the rest of the residues sometimes forming short helices. Our predictions for the location and length of segments with secondary structure are in reasonabl agreement with those from the PHD program (Fig. 2).

Given these pr dictions, we r examined the alignm nt of the sequences of POR, OmpF, and GLUT1. We

verified that segments known t have secondary structure in one or both porins aligned well with segments for which we predicted secondary structure in GLUT1 (Fig. 2). Eleven of the 16 predicted β-strands in GLUT1 overlapped partially with β-strands in porins. The paucity of gaps in these regions with conserved secondary structure is noteworthy. Some of the remaining β-strands in the porins correspond to segments predicted as helices in GLUT1 and vice versa. The alignment in Fig. 2 comprised about the last 400 residues in GLUT1; based on additional alignments, the N-terminal region of GLUT1 might have originated in partial duplication of a porin gene. In addition, there is a high degree of sequence conservation among members of the GLUT family, and hence a multi β-strand motif may be applicable to all of them.

Three-dimensional Model of the βB in GLUT1. The predictions above suggested to us that GLUT1 might fold as the porins, forming a βB. To visualize whether such an idealized construct was compatible with GLUT function, we built a three-dimensional model of the putative GLUT1 βB, with the more hydrophilic sides of the tm β-strands facing the barrel pore. To ensure that there were no bad Van der Waals contacts, limited energy minimization was performed (300 iterations, conjugate descent algorithm. DISCOVER program). Fig. 5 shows an end-view photograph of the barrel (from inside the cell) including β-D-glucopyranose in its lumen. The Van der Waals inside diameter of the barrel, while irregular, was at least 11A which is more than enough to allow hydrated hexoses to pass through the channel.

Pri r vidence consistent with a βB f ld. *Th 2-N-[4-(1-azi-2, 2, 2-trifluoroethyl)benzoyl]-1,3-bis-(D-mannos-4-yloxy)-2-propylamine (ATB-BMPA) binding site.* Peptide 217-272 appears intracellular, since a sp cific Ab binds to it only when the cell membrane is

perm abilized (Davis et al., 1990, Biochem. J., 266:799-808) This segment is very hydrophilic so that the more hydrophobic tm segment that follows it is likely to begin only at or near residue 273 (in either the 12H, βB, or PHD predictions: Fig. 2). The next marker along the chain is residue 282, which has been recently placed extracellularly, since mutation of it (Gln→Leu) decreases ATB-BMPA exofacial binding by 95% (Hashiramoto et al., 1992, J. Biol. Chem., 267:17502-17507). Hence, segment 273-281 likely spans the membrane: this segment (9 residues) is too short to be a tm α-helix (Chin, et al., 1987, J. Biol. Chem., 261:7101-7104) residues but has the correct length for a tm β-strand (strand 9, residues 271-280, Figs. 4 and 6). In the 12H model, residue 282 was placed at the center of tm α-helix 7, where it would be inaccessible to ATB-BMPA. In the βB model, residue 282 is instead in an extracellular connecting loop.

The proportions of α and β structures in GLUT based on CD and FTIR spectroscopy. This issue is unsettled. From FTIR spectroscopic evidence, it was concluded that GLUT1 displays distinct vibrations for α-helical structure while those for β-structure are absent (Chin, supra.). This was partly challenged by a later FTIR study, which also found GLUT1 to be predominantly α-helical but in addition found evidence strongly suggesting the presence of some β-structure, with a portion of it forming antiparallel strands (Alvarez et al., 1987, J. Biol. Chem., 262:3502-3509). Interpretations of CD evidence also appear divided. In one case, CD was said to indicate the presence in GLUT1 of som 82% α-helices, 10% β-turns and 8% random structure, with no β-strands. (Chin et al., 1987, Proc. Natl. Acad. Sci. USA, 84:4113-4116). However, mor recently, use of an algorithm (Perezel et al., 1991, Protein Eng., 4:669-679) to analyze CD data led

to predictions (Park et al., 1992, Protein Sci., 1:1032-1049) of β-structure in GLUT1, POR and OmpF, among other membrane proteins. Our assignments for GLUT1 structure are in line with the more recent FTIR and CD studes (Alvarez et al., supra.; Park et al., supra.).

Solvent accessibility of the GLUT backbone is better explained by the βB model. Others and ourselves have reported evidence for the existence of a water-filled pore across GLUTs (Alvarez et al., 1987, J. Biol., Chem., 262:3502-3509; Jung et al., 1986, J. Biol. Chem., 261:9155-9160; Fischbarg et al., 1990, Proc. Natl. Acad. Sci. USA., 87:3244-3247). Such an open pathway would have to coexist with an apparent enzyme-type tight-fitting structure, since GLUTs display steric selectivity for substrates. This apparent contradiction may be resolved by noting that the water permeability of GLUTs (Fischbarg et al., 1993, Alfred Benzon Symp., 34:432-446) is only some 7% that of specific water channels (Preston et al., 1992, Science, 256:385-387), as if water traverses an open pathway through GLUTs only during part of a cycle of conformational changes. Both the 12H and βB models imply a hydrophilic pore in GLUT. On the basis of hydrogen-deuterium exchange, however, ≈90% of the GLUT1 amine protons are exchanged almost immediately (Alvarez et al., supra.; Hans et al., 1992, Trends Biochem. Sci., 17:328-333). These exchange data can be explained more readily if GLUT1 is a βB with a solvent-filled pore, as in that case most backbone amine hydrogens lining the pore and forming connecting loops would be accessible to solvent.

GLUT1 as a multifunctional βB transporter. From recent evidence, compounds other than sugars such as water (Fischbarg t al., 1990, Proc. Natl. Acad. Sci. U.S.A., 87:3244-3247; Zhang et al., 1991, J. Clin.

Invest., 88:1553-1558), nicotinamide (S fue et al., 1992, Biochem. J., 288:669-674), and dehydroasc rbic acid (Vera et al., 1993, Nature, 364:79-82) traverse GLUTs, suggesting that GLUTs are multifunctional (Sofue et al., supra.). Since a barrel framework is essentially fixed, as argued for porins, the GLUT1 connecting loops might operate as molecular gates and might also be involved in binding solutes and discriminating among them. The putative long intracellular GLUT1 loop (residues 204-270) may be an example, since glucose binding to the loop induces a conformational change in it (Asano et al., 1992, FEBS. Lett., 298:129-132) and antibodies against the peptide Asn-217-I13-272 inhibit the binding of cytochalasin B to the protein). This loop may also have a binding site for ATP (Lys-225-Lys-229) (Carruthers et al., 1989, Biochemistry, 28:8337-8346) and protein kinase C phosphorylation sites (Ser-226, Ser-24[illegible]) (Deziel et al., 1989, Int. J. Biochem., 21:807-814), all with potential functional roles. Lastly, all three antibodies we tested bind to putative mobile loops and enhance DOG uptake. The topology we propose is summarized in Fig. 6.

7. Example: Further Proteins Shown To Include Beta-barrel Structure

7.1. Materials And Methods

We obtained from databases (Swissprot, Protein Information Resource) the sequences of:

sw:p06009 Reaction center protein L chain (RCL).

sw:p02945 Bacteriorhodopsin precursor (BR)

sw:p04480 Colicin A (COLA).

pir3:s16070, *Rhodobact r capsulatus* porin (POR; SEQ ID NO:2).

sw:p02931 *Escherichia coli* porin (Ompf; SEQ ID NO:1).

sw:p11166 Glucose transporter type 1 (GLUT1; SEQ ID NO:3), erythrocyte/brain.

sw:p29972 Water channel protein for red blood cells and kidney proximal tubule (CHIP28).

sw:p02710 Acetylcholine receptor protein, alpha chain precursor

sw:p02920 Lactose permease

sw:p13866 Sodium/glucose cotransporter

sw:p08513 Potassium channel protein, larval (shaker-epsilon)

sw:p16614 Calcium-transporting ATPase sarcoplasmic reticulum type

sw:p20648 Potassium-transporting ATPase alpha chain (proton pump, gastric H^+/K^+-ATPase).

(accession codes are given in parenthesis).

Predictions. Several algorithms were used. For hydropathy analysis, we calculated the average hydrophobicity H_i for a span of i residues using the Kyte-Doolittle (KD) scale (Kyte and Doolittle, 1982, J. Mol. Biol., 157:105-132). We used spans of 21 and 7 residues. A span of 21 residues is appropriate because membrane spanning α-helices are of this or similar lengths. On the other hand, a shorter span can uncover trends in the hydrophobicity profile that the larger span might average out. We decided on 7 residues as the shortest span to give a representative picture of a local neighborhood in a chain without giving rise to excessive "noise". We also used the Union algorithm, described above, to predict protein segments expected to be transmembrane, namely, having relatively high hydrophobicity and propensity to form amphiphilic α or β structures.

We also employed the Ch u-Fasman predicit n method as implemented in the Chou-Fasman-Prevelig (CFP) algorithm (Prevelige et al., "Ch u-Fasman prediction of

the secondary structur of pr teins: Chou-Fasman-Prevelige algorithm. In: G.D. Fasman (eds). Prediction of protein structure and the principles of protein conformation", Plenum Press, New York, New York, pp. 391-416 (1989)). Our figures showed α and β average propensities calculated for tetrapeptides and assigned to the first residue, following the CFP procedure. Where these propensities equal or surpass the CF threshold (100 in their units), we mark the segments (α prd and β prd lines). We also show the CF <pt> propensity; where <pt> exceeds the threshold recommended in the CFP procedure (0.00075), 4-residue predicted turns are marked by lines (denoted as "t prd") beginning with the suprathreshold residue. Our routine simply marks all such 4-point turns, rather than attempting to opt between them (as in the CFP procedure) when they overlap.

We also used the results obtained with the PHD neural network prediction program (Rost and Sander, 1992, Nature, 360:540), which runs without human intervention in a computer, and is therefore unbiased to that extent.

We found that, as a rule, no single procedure was completely sufficient, and it was best to combine in one figure several different types of plots so as to compare them and derive a global picture for a given protein. To that end, we wrote a program ("UCFP") in the PowerBasic language (Power-BASIC Inc., Brentwood, Ca 94513), compiled it, and ran the executable file under IBMDOS. The source code of UCFP is set forth in Section 7.1.1, below. UCFP is a predecessor of the UNION program, and uses as inputs two files: a) the amino acid sequence of a protein, and b) a file with literal secondary structural assignment codes for that s quence, either taken from the Brookhav n database for proteins with known structure or derived from

predicti ns f r proteins of unknown structure. Our program computes hydrophobicities, U and CFP α, β and pt propensities, converts the literal structure codes into numbers, and generates a columnar output file. We obtained the figures presented here by importing UCF output into a graphics program ("Origin", MicroCal Software, Northampton, MA 01060). We also found useful the graphic display program "PSAAM" (Crofts, AR, "Protein Sequence Analysis and Modeling for Windows 3 []," University of Illinois, Urbana, IL (Ph.D.; Dissertation)) to verify the validity of our algorithms.

```
53      if nseq=0 then nseq=1999
54      naa = 20      'unless dealing with extraterrestrials...
55      'NSEQ = 300  'maximum number of amino acids in sequenc ; sets array sizes
56      ' program sensitive to this in the Power Basic environment
57      '   however, if compiled, so far no limit encountered for the executable
58      DIM symbols$(naa)
59      DIM seqh(NSEQ)
60      DIM seqn(NSEQ)
61      DIM seq$(NSEQ)
62      DIM vkyte(naa)
63      DIM pturn(naa)
64      DIM hout(NSEQ)
65      DIM houtsh(NSEQ)
66      DIM amphi(nseq)
67      DIM amphiout(NSEQ)
68      DIM amphibeta(NSEQ)
69      DIM amphialpha(NSEQ)
70      DIM UALPHA(NSEQ)
71      DIM U(nseq)
72      DIM ubeta(NSEQ)
73      DIM DUM(NSEQ)
74      DIM DUMN(NSEQ)
75      DIM pai(naa)          'alpha propens. for individual amino acids
76      DIM pas(nseq)         'sequential indiv. alpha propens. along chain
77      DIM patetr(nseq)
78      DIM pam$(nseq)
79      DIM pbi(naa)          'beta propens. for individual amino acids
80      DIM pbs(nseq)         'sequential indiv. beta propens. along chain
81      DIM pbtetr(nseq)
82      DIM pbm$(nseq)
83      DIM pt(naa, 4)
84      DIM PTSEQ(NSEQ)
85      DIM ptm$(nseq)
86      DIM phda$(nseq)
87      DIM phdb$(nseq)
88      DIM phdt$(nseq)
89      DIM temp$(nseq)
90      comienzo:
91      aacodes$ = "ARNDCQEGHILKMFPSTWYV"
92      alphacut = 100
93      betacut = 100
94      turncut = 0.75e-4
95      CLS
96      drive$ = "c:"
97      path$ = "\UCFP"
98      DEFAULT$ = drive$ + path$ + "\" + "UCFP.INI"
99      OPEN DEFAULT$ FOR INPUT AS #5
100     INPUT #5, drive$,path$,filename$,filename$,angalpha,angbeta,WHEL,span
101     CLOSE #5
102     fileinp$ = drive$ + path$ + "\" + filename$ + ".sqt"
103     fileout$ = drive$ + path$ + "\" + filename$ + ".dat"
104     '*****************************************************************************
105     'Kyte-Doolittle scale
```

```
106     '      1,  2,   3,   4,  5,   6,   7,   8,   9,  10
107     '      A,  R,   N,   D,  C,   Q,   E,   G,   H,   I
108      DATA 1.8,-4.5,-3.5,-3.5,2.5,-3.5,-3.5,-0.4,-3.2,4.5
109     '      11,  12, 13, 14,  15,  16,  17,  18,  19, 20
110     '       L,   K,  M,  F,   P,   S,   T,   W,  Y,   V
111      DATA  3.8,-3.9,1.9,2.8,-1.6,-0.8,-0.7,-0.9,-1.3,4.2
112     '*************************************************************************
113     'CHOU-FASMAN 64-protein database
114     '*************************************************************************
115     DATA A,139, 79,0.060,0.076,0.035,0.058
116     DATA R,100, 94,0.070,0.106,0.099,0.085
117     DATA N, 78, 66,0.161,0.083,0.191,0.091
118     DATA D,106, 66,0.147,0.110,0.179,0.081
119     DATA C, 95,107,0.149,0.053,0.117,0.128
120     DATA Q,112,100,0.074,0.098,0.037,0.098
121     DATA E,144, 51,0.056,0.060,0.077,0.064
122     DATA G, 64, 87,0.102,0.085,0.190,0.152
123     DATA H,112, 83,0.140,0.047,0.093,0.054
124     DATA I, 99,157,0.043,0.034,0.013,0.056
125     DATA L,130,117,0.061,0.025,0.036,0.070
126     DATA K,121, 73,0.055,0.115,0.072,0.095
127     DATA M,132,101,0.068,0.082,0.014,0.055
128     DATA F,111,123,0.059,0.041,0.065,0.065
129     DATA P, 55, 62,0.102,0.301,0.034,0.068
130     DATA S, 72, 94,0.120,0.139,0.125,0.106
131     DATA T, 78,133,0.086,0.108,0.065,0.079
132     DATA W,103,124,0.077,0.013,0.064,0.167
133     DATA Y, 73,131,0.082,0.065,0.114,0.125
134     DATA V, 97,144,0.062,0.048,0.028,0.053
135
136     ' ////////////////////////////////////////////////////////////
137     CLS
138     'print "Free memory: ";fre(0); fre(-1); fre(-2)
139
140     PRINT "UCFP ALGORITHM; J. Fischberg, F. Czegledy, P. Iserovich. Copyright 1994"
141     print" Set for sequence lengths up to  " nseq
142     print "
143     START:
144     DUMMY = 0
145     PRINT " ENTER ONE OF THE FOLLOWING "
146     PRINT
147     PRINT "1. CHANGE FILE NAME FOR INPUT; currently: "; fileinp$
148     'PRINT "2. CHANGE FILE NAME FOR OUTPUT; currently: "; fileout$
149     PRINT "2. CHANGE ANGLE FOR ALPHA/BETA STRUCTURES (HYDROPHOBIC MOMENT) "
150     PRINT "   ALPHA STRUCTURE= "; angalpha; "BETA STRUCTURE= "; angbeta
151     PRINT "3. CHANGE A.A. SPAN FOR MEMBRANE HELICES; currently: "; WHEL
152     PRINT "4. CHANGE A.A. SPAN FOR UNION; currently: "; span
153     PRINT "5. CHANGE PATH; CURRENTLY: " path$
154     print "6. CHANGE DRIVE; CURRENTLY: " drive$
155     'PRINT "6. CHANGE WINDOW SIZE FOR SMOOTHING UNION; currently: "; UW
156     PRINT "9. TO END SESSION WITHOUT RUNNING"
157     PRINT "0. (DEF=CR) MAIN - RUN WITH CURRENT PARAMETERS- RUNS ONLY ONCE AND EXITS
158     print
159     PRINT " "
160      INPUT DUMMY
```

```
161       SELECT CASE DUMMY
162       CASE 1 :   GOSUB FILENAMEINPUT
163      ' CASE  2 :  GOSUB FILENAMEOUTPUT
164       CASE 2  :  GOSUB MOMENTANGLE
165       CASE 3  :  GOSUB ALPHAWINDOW
166       CASE 4  :  GOSUB BETAWINDOW
167       CASE 5  :  GOSUB NEWPATH
168       'TO COMIENZO   'read under repairs- monkeying with discouraged
169      ' CASE 6  : GOSUB UNNWINDOW
170       CASE 9  :  GOTO salida
171       CASE 0  :  GOTO correte
172     END SELECT
173     GOTO start
174
175     '////////////////////////////////////////////////////////////////////////////
176     newpath:
177     print "Enter new path, e.g., C:\PROT\PRD(note no end \ or end space)"
178     input test$
179     if test$="" then goto newpath
180     path$ = test$
181     DEFAULT$ = drive$ + path$ + "\" + "UCFP.INI"
182     fileinp$ = drive$ + path$ + "\" + filename$ + ".sqt"
183     fileout$ = drive$ + path$ + "\" + filename$ + ".dat"
184     return
185     '////////////////////////////////////////////////////////////////////////////
186     correte:  'main routine - records parameters and runs
187       cls
188       OPEN DEFAULT$ FOR OUTPUT AS #6
189       WRITE #6,drive$,path$,filename$,filename$,angalpha,angbeta,WHEL,span
190       CLOSE #6
191       GOSUB WORKING
192     PRINT
193     print "COMPLETED RUN SUCCESSFULLY - STOPPING NOW"
194     STOP  'With all the erase statements to save memory, cannot run again
195       'all key arrays erased by now
196     'GOTO start
197     '////////////////////////////////////////////////////////////////////////////
198
199     FILENAMEINPUT:
200     cls
201     'print "string space remain." fre(0)
202     'print "bytes left in mem. f. data"  fre(-1)
203     'print "stack space never used " fre(-2)
204     'input dummy
205     chdrive drive$
206     chdir path$
207     files  "*.SQT"
208     rem SHELL "DIR " + path$ + "*.SEQ/W"
209     LOCATE 20, 1
210     PRINT "ENTER FNAME (ONLY) FOR INPUT SEQ; PROG ADDS DEF
          FILE TYPES .SQT & .DAT"
211     INPUT filename$
```

```
212     fileinp$ = drive$ + path$ + "\" + filename$ + ".sqt"
213     fileout$ = drive$ + path$ + "\" + filename$ + ".dat"
214      RETURN
215
216     MOMENTANGLE:
217     PRINT "ENTER ANGLE FOR ALPHA STRUCTURES "
218     INPUT angalpha
219     PRINT "ENTER ANGLE FOR BETA STRUCTURES "
220     INPUT angbeta
221       RETURN
222
223     ALPHAWINDOW:
224     PRINT "ENTER WINDOW SIZE FOR  MEMBRANE HELICAL SPANS (ODD NUMBER)"
225     INPUT WHEL
226       RETURN
227
228     BETAWINDOW:
229     PRINT "ENTER WINDOW SIZE FOR  UNION SPAN (ODD NUMBER)"
230     INPUT span
231       RETURN
232
233     UNNWINDOW:
234      PRINT "ENTER WINDOW SIZE FOR SMOOTHING UNION"
235     INPUT UW
236       RETURN
237
238     ' ////////////////////////////////////////////////////////////////////////
239
240     WORKING:
241     print fre(0); fre(-1); fre(-2)
242     OPEN fileinp$ FOR INPUT AS #1
243     INPUT #1, sequence$
244     input #1, structure$
245     CLOSE #1
246     FOR n = 1 TO 20
247       READ vkyte(n)
248       symbols$(n) = MID$(aacodes$, n, 1)
249     NEXT n
250     FOR i = 1 TO 20
251     READ symbols$(i),pai(i),pbi(i),pt(i,1),pt(i,2),pt(i,3),pt(i,4)
252     NEXT i
253     RESTORE
254          '///////  ************** ////////////////////////
255     PRINT "..............  WORKING.................."
256     cfspan = 4       'prepare for Chou-Fasman-Prevelige tetrapeptides
257     nlen = LEN(sequence$)
258     FOR n = 1 TO nlen
259     seq$(n) = MID$(sequence$, n, 1)  'list of aa codes
260     NEXT n
261     FOR l = 1 TO nlen       ' from 1 to length of sequence */
262       FOR k = 1 TO 20
263       IF seq$(l) = symbols$(k) THEN ' identify ordinal FOR aa */
264         seqh(l) = vkyte(k)  '  assign hydrophobicity value to residue*/
265         seqn(l) = k          'assign  residue name number /
266             pas(i) = pai(k)   'assign alpha propensity
267             pbs(i) = pbi(k)   'assign beta propensity
268             exit for       'done h re; leave f r/next loop
```

```
269       END IF
270         NEXT k
271      NEXT I
272
273      FOR n = 2 TO (nlen - 2)
274      PTSEQ(n)=pt(seqn(n-1),1)*pt(seqn(n),2)*pt(seqn(n+1),3)*pt(seqn(n+2),4)
275      NEXT n
276       erase seqn
277      PTSEQ(1) = PTSEQ(2)
278      PTSEQ(nlen) = PTSEQ(nlen - 2)
279      PTSEQ(nlen - 1) = PTSEQ(nlen - 2)
280      for i=1 to nlen
281      if ptseq(i) >= turncut then
282          for ind = 0 to 3
283        ptm$(i + ind) = "4.5 " : next ind : goto cortada
284      end if
285      if ptseq(i) < turncut then
286        if ptm$(i) =  "4.5 " then goto cortada
287        else
288          ptm$(i) = "    "
289      end if
290      cortada:
291      next i
292      CALL NORMAL(PTSEQ(), nlen, ptseq())
293
294      ' ************ ////////////////// **************
295
296        ' HYDROPHOBICITY CALCULATION FOR MEMBRANE HELICES * /
297      FLAG = 1 ' calculate hydrophobicity* /
298      j = WHEL                ' window
299      GOSUB MAIN        ' and we will get hout(m)
300      CALL NORMAL(hout(), nlen, hout()) 'and we will get hout long*/
301
302      ' ************ ////////////////// **************
303
304        ' HYDROPHOBICITY CALCULATION FOR SHORT SPAN
305      FLAG = 2 ' calculate hydrophobicity* /
306      j = span                ' window
307      GOSUB MAIN        ' and we will get houtsh(m)
308      CALL NORMAL(houtsh(), nlen, houtsh()) 'and we will get hout short*/
309      ' ************ ////////////////// **************
310      ' CALCULATION OF TETRAPEPTIDE PROPENSITIES
311      j = cfspan
312      for i=1 to nlen-3
313      patetr(i) = ( pas(i) + pas(i+1) + pas(i+2) + pas(i+3) )/cfspan
314      pbtetr(i) = ( pbs(i) + pbs(i+1) + pbs(i+2) + pbs(i+3) )/cfspan
315      if patetr(i) >= alphacut then
316        pam$(i) = "4.5 "
317      else
318        pam$(i) = "    "
319      end if
320      if pbtetr(i) >= betacut then
```

```
321       pbm$(i) = "4.5 "
322     else
323       pbm$(i) = "    "
324     end if
325     next i
326      erase pas, pbs
327     for j= 2 to 0 step -1     'approximate bottom ends
328     patetr(nlen-j) = patetr(nlen-3)
329     pbtetr(nlen-j) = pbtetr(nlen-3)
330     next j
331
332     CALL NORMALPA(patetr(), nlen, patetr() )
333     CALL NORMALPB(pbtetr(), nlen, pbtetr() )
334
335         ' calculation alpha moment*/
336
337     FLAG = 0  'selects amphiout output
338     j = span
339     ANGLE = angalpha
340     GOSUB MAIN           'gets amphiout(m)*/
341     CALL NORMAL(amphiout(), nlen, amphialpha())'gets amphialpha*/
342
343     ' ************ ////////////////// **************
344     ' calculation beta moment*/
345
346     FLAG = 0 'selects amphiout output
347     j = span
348     ANGLE = angbeta
349     GOSUB MAIN      'gets amphiout(m)*/
350     CALL NORMAL(amphiout(), nlen, amphibeta())  'gets amphibeta
351      erase amphiout
352     ' ************ ////////////////// **************
353
354     'calculate union alpha
355     CALL UN(ptseq(), houtsh(), amphialpha(), nlen, ualpha())
356      erase amphialpha
357     CALL NORMAL(ualpha(), nlen, ualpha())
358     'calculate union beta
359     CALL UN(ptseq(), houtsh(), amphibeta(), nlen, ubeta())
360      erase amphibeta
361     CALL NORMAL(ubeta(), nlen, ubeta())
362     erase amphi
363     erase seqh
364     ' ************ ////////////////// **************
365     ' ////////////////////////////////////////////////////////////////
366     ' PROCESS STRUCTURE STRING (PREDICTIONS OR CRYSTALLOG.)
367     alfam$ = "3.5 "
368     betam$ = "3.5 "
369     turnm$ = "3.5 "
370     FOR n = 1 TO nlen
371     temp$(n) = MID$(structure$, n, 1)  'list of structure codes
372     NEXT n
373     for i=1 to nlen
374     if temp$(i) = "H" then
375       phda$(i) = alfam$  :  phdb$(i) = "    " : phdt$(i) = "    "
376     end if
377     if temp$(i) = "E" then
```

```
      phda$(i) = "    "  :  phdb$(i) = beta$ : phdt$(i) = "    "
    end if
    if temp$(i) = "C" then
      phda$(i) = "    "  :  phdb$(i) = "    "  : phdt$(i) = "    "
    end if
    if temp$(i) = "T" then
      phda$(i) = "    "  :  phdb$(i) = "    "  : phdt$(i) = turn$
    end if
    next i
    close #4
    erase temp$
    GOsub producto
    return

    ' ////////////////////////////////////////////////////////////////////////

    MAIN:       'window size j is already defined
    IF j > 1 THEN
    '*************** STARTING SEGMENT
     FOR m = 1 TO (j - 1) / 2       '1 to 10
     LB = 1          'LOW BOUNDARY
     UB = m + (j - 1) / 2'upper BOUNDARY  m+10
     GOSUB CALCULATION
     NEXT m
    ' ********MAIN CENTER SEGMENT     '11 to nlen - 10
     FOR m = (j + 1) / 2 TO (nlen - (j - 1) / 2)' m  center of the window*/
     LB = m - (j - 1) / 2   ' m-10
     UB = m + (j - 1) / 2   ' m+10
     GOSUB CALCULATION
     NEXT m
    '******** END SEGMENT 'nlen-9 to nlen
     FOR m = (1 + nlen - (j - 1) / 2) TO nlen 'm  ctr of window*/
     LB = m - (j - 1) / 2 'low BOUNDARY  m-10
     UB = nlen
     GOSUB CALCULATION
     NEXT m
   END IF
     RETURN

    ' ////////////////////////////////////////////////////////////////////////

    CALCULATION:

    IF FLAG = 1 THEN  ' calculate hydrophobicity of std. tm. segmts.*/
    cumh = 0: cum = 0              ' reset hydrophobicity accumulators */

    FOR I = LB TO UB             'loop on i through all res. in window */
    cumh = seqh(I) + cumh
    cum = cum + 1
    hout(m) = cumh / cum             ' compute hydrophobicity average *
    NEXT I

    ELSEIF FLAG = 2 THEN  ' calculate hydrophobicity of short tm. segmts.*/
    cumh = 0: cum = 0              ' reset hydrophobicity accumulators */

    FOR I = LB TO UB             'loop on i through all res. in window  /
    cumh = seqh(I) + cumh
    cum = cum + 1
```

```
436     houtsh(m) = cumh / cum                  ' compute hydrophobicity average *
437     NEXT I
438
439     ELSEIF FLAG = 0 THEN                    ' calc. hydrophobic moment*/
440
441     t = 0 : acum = 0 : Mx = 0! : My = 0!     'reset amphi accumulators /
442     FOR I = LB TO UB                        'loop on i through all res. in window */
443     x = COS(2 * 3.1416 * ANGLE * (I - LB) / 360)  'Eisenberg
444     y = SIN(2 * 3.1416 * ANGLE * (I - LB) / 360)
445     Mx = Mx + (x * seqh(I))
446     My = My + (y * seqh(I))
447     acum = acum + 1
448     NEXT I
449     amphiout(m) = SQR(Mx ^ 2 + My ^ 2)
450
451     END IF
452
453       RETURN
454
455     ' ////////////////////////////////////////////////////////////////
456
457     producto:
458     c$ = " , "
459     OPEN fileout$ FOR OUTPUT AS #2
460     PRINT #2, "res" c$ "aa" c$ "H21" c$ "H7" c$ "ua" c$ "ub" c$ "Pa" c$_
461     "am" c$ "Pb" c$ "bm" c$ "pt" c$ "tm" c$ "prda" c$ "prdb" c$ "prdt"
462     FOR l = 1 TO nlen
463     locate 16,1
464     PRINT "l= "; l; "; nlen= "; nlen
465       hlng = round(hout(l), 2)
466       hsh = round(houtsh(l), 2)
467       ua = round(ualpha(l), 2)
468       ub = round(ubeta(l), 2)
469       pa = round(patetr(l), 2)
470       pb = round(pbtetr(l), 2)
471       pt = round(ptseq(l), 2)
472     PRINT #2,l c$ seq$(l) c$ hlng c$ hsh c$ ua c$ ub c$ pa c$ pam$(l)_
473      c$ pb c$ pbm$(l) c$ pt c$ ptm$(l) c$ phda$(l) c$ phdb$(l) c$ phdt$(l)
474     NEXT l                              'and do next row
475     CLOSE #2
476     erase hout,houtsh,ualpha,ubeta,patetr,pbtetr,ptseq
477     return
478
479     ' ////////////*****************************/////////////////
480
481     salida:
482     PRINT " ### DONE ####"
483
484     stop
485     end
486
487
488     ' ////////////////////////////////////////////////////////////////////////////
489
490           SUB NORMAL (DUM(), nlen, DUMN())
491
492     ytop#=DUM(1):ybot#=DUM(1):yCUM#=DUM(1)' reset max, min & avg accumulators*/
493       FOR I = 2 TO nlen 'MAX AND MIN DETERMINATION  ' /
```

```
494         IF DUM(I) > ytop# THEN
495                         ytop# = DUM(I)
496                         yhord = I
497                         end if
498         IF DUM(I) < ybot# THEN
499                         ybot# = DUM(I)
500                         ylord=I
501                         end if
502       '   yCUM# = yCUM# + DUM(I)
503         NEXT I
504       '   yAVeraG = yCUM# / nlen 'average
505         FOR I = 1 TO nlen
506       'print DUM(I)
507       'print (DUM(I) - ybot#);(ytop# - ybot#);(DUM(I) - ybot#)/ (ytop# - ybot#)
508         DUMN(I) = -4.5 + 9 * (DUM(I) - ybot#) / (ytop# - ybot#)
509         NEXT I
510
511         END SUB
512
513       ' ////////////////////////////////////////////////////////////////////////////
514
515         SUB UN (ptseq(), HOUTsh(), AMPHI(), nlen, U())
516
517       FOR m = 1 TO nlen
518         U(m) = HOUTsh(m) + AMPHI(m) - ptseq(m)
519       NEXT m
520         END SUB
521       ' ***************************************************
522             SUB NORMALPA (DUM(), nlen, DUMN())
523
524       deltalfa# = 75
525       llalfa# = 64
526           FOR I = 1 TO NLEN
527         DUMN(I) = -4.5 + 9 * (DUM(I) - llalfa#) / (deltalfa#)
528         NEXT I
529
530         END SUB
531       ' ***************************************************
532             SUB NORMALPB (DUM(), nlen, DUMN())
533
534       deltabeta# = 106
535       llbeta# = 51
536           FOR I = 1 TO NLEN
537         DUMN(I) = -4.5 + 9 * (DUM(I) - llbeta#) / (deltabeta#)
538         NEXT I
539
540         END SUB
```

7.2. Results

Evaluation of the prediciton profiles.

Validations: (1) multihelical proteins. The reaction center (L chain) constitutes a good example of a successful prediction (Fig. 7a). The H7 profile marks several hydrophobic segments which are long enough to span the membrane. This prediction of long segments is borne out by the $U_{\alpha 7}$ peaks, and by the relative paucity of predicted turns, leaving long stretches of sequence with little turn propensity and hence with relatively higher propensity to form structure. An assignment to multi-α folding could be made at this point, after which the segments could be refined using a detailed CFP spreadsheet (Prevelige, supra.), cap propensities, and so on.

Bacteriorhodopsin also evidences long hydrophobic stretches (H_7) borne out by $U_{\alpha 7}$ peaks, and relatively few predicted turn regions (Fig. 7b). The trend to long structured segments is curiously more discernible in the CFP-β predictions than in the α-predictions. Still, the protein can be classed as multi-α on the basis of the length of the predicted segments.

For colicin (Fig. 7c), hydrophobicity analysis alone seems insufficient, since it predicts long stretches known as transmembrane as hydrophilic. Our way of plotting normalized (rather than absolute) H values exaggerates this trend, which is nonetheless noteworthy. In this instance, CFP α-predictions and $U_{\alpha 7}$ profiles demonstrate that the length of the predicted segments is consistent with multi-α-helical fold. Overall CFP α-propensity is higher than that for β (we plot absolute values for both). Hence, multi-α assignment seems adequate.

Validati ns: (2) porins. *Rhodobact r capsulatus* porin exemplifies a trend (Fig. 7d): some p aks that appear

as long hydrophobic stretches in the H_{21} profile are split in the H_7 profile. Even if qualitative judgments are tentative, this does not happen to the same extent in RCL or BR. Neither the α nor the β predictions mark long segments here, and turn propensity peaks appear frequently along the chain. A tentative assignment of multi-β fold can be made at this point. If one then focuses attention on the U_{b7} peaks, one can verify that they mark the β-segments exceedingly well. At a threshold of 1.83, all strands will be marked, with minimal overprediction. Segments lengths could be further refined as above.

The *Escherichia coli* porin profiles (Fig. 7e) show further the limitations of hydrophobicity analysis *per se*. The hydrophobicity profiles largely miss the β-hairpin between residues 35-65. However, the CFP β-predictions and U for β segments find them. The U peaks are especially noteworthy; as above, with a threshold of 2.15, U marks all the β-strands with minimal overprediction. One can note also repeated suprathreshold turn predictions, seemingly at regular intervals; from all this, a tentative assignment of multi-β structure may be made. This plot also allows the rare opportunity of evaluating the performance of the PHD robot by comparing the structure derived from crystallography with a prediction PHD made of this protein shortly before it was incorporated to its database. Practically all structured segments are detected by PHD, which also does reasonably well in predicting their lengths. Once more, those lengths are too short for transmembrane α-helices, but adequate for β-strands, confirming the tentative assignment above. There is another feature of the PHD prediction worth noting: as many as 7 β-s gments are predicted as α-helical, while one of the short α-helic s is predicted as a β-strand. Such types of mispredicti ns can be

common when humans mak their own judgments, so here the computer brings no improvement. The advantage a human has is to know that the protein is in a membrane, and hence that the structured segments predicted as α-helical are too short to be transmembrane, pointing instead to a β-barrel.

In closing this section, we note that the group of proteins reviewed so far has a common feature: they tend to have relatively short sequences, not exceeding some 350 amino acids. Perhaps that has made crystallizing them somewhat easier, certainly predictions also appear relatively straightforward, compared with some for the longer sequences.

Proteins with unknown structure.

The prediction profiles for facilitative glucose transporter indicate a number of short segments (Fig. 8a). PHD predicts only three long segments as α-helical. Yet, of these, the middle one (#230-260) forms part of a known long intracellular loop, and might be actually broken by a turn. In view of the number of predicted short segments, the remaining two long segments could not suffice to label the protein as multi-α-helical. When analyzed more closely, predicted turns can be discerned that could interrupt those segments. In contrast, the U_{b7} peaks, β predictions, and a good number of PHD-predicted segments are in register and give a cogent picture of short segments, the approximate location of which we mark with arrows. One can note how the segments may be nested between predicted turns. Partly on this basis, we have predicted for this protein a porin-like multi-β folding, with some α-helic s in the connecting loops. W also show in the fourth panel the p ssible orientation of the predicted β-strands.

CHIP28: PHD predicts (Fig. 8b) only two segments long enough to be transmembrane as α-helices, which makes assignment as multi-helical somewhat doubtful. In addition, turn propensities are rather high and repeated along the chain, which speaks for short structured segments. The next feature to are the β-predictions, $U_{\beta 7}$ peaks and PHD-predicted β segments in register along the second half of the sequence. If one now returns to the two long segments, predicted turns are discernible that could break them (one supra, one sub-threshold). Therefore, we assign the protein as multi-β, and mark the 16 putative segments that would give it a porin-type fold. In this view, given its short sequence there would be little in this structure aside from the barrel itself, since the connecting loops would be rather short (except perhaps for the segment 110-140). One might think of it as a rudimentary or bare-bones channel protein.

The acetylcholine receptor α subunit: The H_{21} profile (Fig. 8c) yields several hydrophobic stretches long enough to be transmembrane α-helices; these (M1-4) have been recognized for years. One of the long stretches (M2) is under particular scrutiny as a firm candidate to line the channel (see Karlin A, 1991, "Explorations of the nicotinic acetylcholine receptor", The Harvey Lecture series 85:71-107) for how the different subunits might join to form a channel.)

On the other hand, in the profiles shown here, detail multiplies as one progresses from H_{21} to the other ones. It seems particularly noteworthy that the CFβ and U_{b7} propensities and PHD segment predictions are in register throughout the sequence. The CFα and Hα7 propensities are not, which gives a tentative indication of multi-β folding. We have marked with arrows some segments as the putative 16 β-strands of a porin fold. In Akabas et al., 1992, Science, 258:307-

310, evidence from cysteine-substituted mutants led the authors to describe segment 248-254 as probably forming a β-strand. Our prediction also finds β-structure in that region.

The potential participation of residues 1-200 in forming a channel has been apparently neglected so far, presumably because in the current view, all five receptor subunits would join together and instead simply form a channel lined by their M2 segments. However, these two views are not necessarily mutually exclusive, as a comparison with porins may show. Porins consist of trimers in which each monomer forms its own channel at one end of the molecule. At the other end, however, the individual channels merge into one large opening for the trimer. One wonders whether other membrane proteins may show a similar arrangement in which channel-containing subunits in varying numbers join in to share their openings. For the acetylcholine receptor, it might explain both the clear evidence for a large opening facing the extracellular space and lined by the 5 monomers (Karlin, supra), and the predictions for the stretch 1-200 if each subunit would form its own channel at their intracellular ends, all channels eventually merging.

Lactose permease. Once more, the CFβ and U_{b7} profiles go in register, especially in the second half of the sequence (Fig. 8d). That cannot be said of the CFα and Uα7 profiles, which again tentatively indicates multi-β folding. PHD predicts some 15 β segments, which reinforces this possibility. PHD also predicts five α-helices with length presumably sufficient to span the membrane (residues 7-26, 72-90, 194-223, 267-287, and 352-376). However, the segment 194-223 is marked by several fusions in the top panel as intracellular, and appears as a hydrophilic region in H_{21}, so it seems logical to discard it. The remaining

four long segm nts, even if helical, do not appear enough to form a transmembrane pore of the dimensions required for lactose permeation. Besides, all of them are potentially interrupted or shortened by suprathreshold turn predictions. In view of all this, multi-β folding appears a more logical choice.

This conclusion goes counter to that drawn in several studies in which evidence for a multi-α-helical fold was presented (Kaback, 1992, Int. Rev. Cytol., 137A:97-125). On the other hand, the possibility of extensive β-folding for the lac permease has been advanced before by Radding (Karlin, supra). More recently, the results of Calamia and Manoil, obtained from fusions, have been cited to support the topology of the 12-helix lac permease model (Kaback, supra) and to support the idea that facilitators conforming to a 6 + 6 hydrophobicity profile are α-helical (Nikaido et al., 1992, Science, 258:936-942). Calamia and Manoil apparently selected the locations of their fusions for the limited aim of discriminating between the 12-helix and the 14-helix lac permease models. The fact that their results support the 12-helix model says little (if anything) about whether a-helical folding is to be favored over an alternative such as the partial β-barrel fold proposed by Radding (Karlin, supra), or over a possible 16-β-strand porin fold. In fact, some of the findings of Calamia and Manoil may be taken as possible indications of β-folding. In their own words, "...it appears that 9-11 apolar membrane spanning segments can suffice to promote efficient alkaline phosphatase translocation across the membrane." Another int rpretation might be that the transmembrane segments referred t would be 9-11 residu s in length, that is, too short to b -helical but quite of the correct length for a transmembran β-strand. In addition, the segment between fusi ns 9 and 10, each

one labeling residues as xtracellular, is long enough that the chain, if consisting of short β-strands, could have entered the cell and returned outside. Lastly, fusion 29 apparently labels a stretch as intracellular when an extracellular location was expected; the small increase in activity of fusion 29 appears of dubious significance in view of the fact that fusion 13, also of small but non-zero activity, may be labeling an intracellular location. In a similar vein, the observed range of alkaline phosphatase activities (as against an ideal all-or-none pattern) poses some question as to which locations the intermediate activities may be labeling. A more substantial link between results of fusions and topology may be clearer if control fusions and subsequent expression can be done with membrane proteins of known structure.

Sodium-glucose cotransporter (Fig. 8e), K^+ channel (Fig. 8f). Several of the patterns already referred to above reappear for these sequences. CFβ, $U_{\beta 7}$ and PHD predictions are in register, while those for CFα and $U_{\alpha 7}$ do not seem to be. Turn potentials rise regularly and delimit segments of 10 residues or less. Once more, a multi-β assignment seems plausible. We have marked with arrows segments that might contribute to porin folds. There is evidence that the functional unit for this K^+ channel is a tetramer (MacKinnon, 1991, Nature, 350:232-235); the comments made above for the acetylcholine receptor apply here as well, namely, each monomer may have its own channel, with all four channels merging into one.

Calcium ATPase (Fig. 2g), H^+/K^+-ATPase (Fig. 8h). The length f the s quences does not allow an intuitive comparison on the same basis as w have done until now. Of course, th se proteins may contain homolog us internal r peats, and if so perhaps a more detailed analysis might be performed on ach rep at. For the

present purposes, we will simply call attention to the CFβ, U_{b7} and PHD predictions in register, the number of comparatively short segments predicted (having the proper length for transmembrane β-strands), and the regularity with which peaks appear in the <pt> profile, all of which is consistent with β-folding.

Profile analysis of environmental similarity. We resorted to this recently developed methodology (Gribskow et al., 1990, "Profile analysis. In: R.F. Doolittle (eds). Methods in Enzymology", Academic Press, New York, pp. 146-159; Bowie et al., 1991, Science, 253:164-170). We chose as terms of comparison two environments, those of RCL and Ompf, and set out to investigate whether membrane proteins of interest would have environmental scores closer to one or the other. The results are summarized in Figs. 9a and 9b. For reasons we discuss below, we think this type of analysis does not perform optimally for membrane proteins. Still, some trends are apparent. The Ompf profile (Fig. 9a) recognizes several porins and members of the major facilitator superfamily of proteins, a group that includes the sugar transporters, and gives them better scores than those of most globular unrelated proteins or BR. Conversely, the RCL profile (Fig. 9b) recognizes the RC M chain and BR better than facilitators or porins.

7.3. Discussion

Translocators: economy of the barrel design. Since a main common function of transporters and channels is to allow passage ("translocation") of solutes across the membrane, in what follows w will refer to them as "translocators". Given a limited number of residues, a β-strand can span a membrane with much fewer of them (beginning with six (Rosenbusch, 1985, EMBO J., 4:1593-1597); 10 is certainly adequate).

Hence, as already noted (Radding, 1991, J. Theor Biol., 150:239-249), much less residues are needed to configure a transmembrane translocation unit if the unit is a β-barrel than if the transmembrane segments are α-helical.

The width of the barrel, and the role of the connecting loops. When contemplating a possible β-barrel model for translocators, it seems logical at first to focus on known β-barrel folds so as to determine which one might have a channel suited for translocation. The choice so far seems limited to two main types, the α-β barrels of isomerase-type enzymes (Farber et al., 1990, TIBS, 15:228-234) and the porins. The β-barrel lumen of the 8-stranded isomerase fold, however, appears to be very small, perhaps only 1-2 angstroms. Of course, the pore of the 16-stranded porins is much wider; in Ompf, even with a loop inside its pore and constricting it, its diameter is 7x11 angstroms (Cowan et al., 1992, Nature, 358:727-733). This is adequate for large solutes, but appears excessive for ionic channels and transporters of small solutes. If such translocators have a porin fold, their pores may be modified by loops. Hence, some connecting loops in translocators may fulfill specific functions such as gating a channel, constricting a channel pore, binding to and hence selecting solutes, binding metabolites and cofactors, signaling destination in protein traffic, etc. Evolutionarily, it seems easier to explain the development of translocators if a common translocation unit was conserved (a 16-stranded β-barrel) and different loops evolved for different functions. A similar scheme was advanced by Nikaido and Saier for bacterial facilitators, except that the transl cati n unit they envisaged was 12-α-helical (Nikaid et al., 1992, Sci nc , 258:936-942). In our view, the common translocation unit would b a β-barrel. With this

proviso, the idea of a c mm n translocation unit could be extended to ionic channels (see Fig. 8f), with suitable loops evolutionarily grafted for each given protein (Nikaido, supra). In fact, a β-barrel model has been previously proposed for the voltage-activated K^+ channel (Bogusz et al., 1992, Protein-Eng., 5(4):285-293).

For an alternative, one would have to consider the evolutionary development of translocators by a process that would have tailored the number of strands and hence the width of the channel to the size of the solute considered. Aside from being overly complex, that is not what the evidence points to for bacterial facilitators (Nikaido, supra). In this light, we deem the work of Radding (Karlin, supra) important to point out the possible presence of β structure in lac permease and the Na^+/H^+ antiporter, a concept with which we agree (cf. our Figs. 9d for the lac permease and 9h for the H^+/K^+-ATPase). On the other hand, the partial β-barrels that he proposes may be more difficult to marry with the evolutionary considerations above. From all this, the porin fold emerges as an interesting candidate for a template common to most if not all translocators.

The connecting loops of barrels. In an anti-parallel β-barrel, the loops connecting one strand with the next one can be relatively short, sometimes no longer than needed for a turn. The arrangement has a certain symmetry in that each strand connects only with the neighboring ones, thereby decreasing potential steric conflicts between different loops. This is what happens in porins. In the view we propose, such loops would be crucial, since the transl cating unit made out of a β-barrel would be too static to result in, say, gating. Conformational changes associated with binding and/or selectivity ar also easier to conceive if they

involve only loops, rather than massive prot in segments.

One might also mention that finite water permeability through proteins has been shown to exist not only across water channels such as CHIP28 or γ-TIP but across several transporters such as GLUT1, the sodium/glucose cotransporter, and CFTR (Hasegawa et al., 1991, Science, 258:1477-1479). Water permeation could of course take place through any type of preferential pathway in a protein, but the presence of β-barrels acting as translocation pathways would provide a ready explanation for water passage through transporters.

Analysis of environmental scores in membrane proteins. The profile analysis methodology has been developed for globular proteins. Hence, in the way it currently stands, the side chains pointing outward from the protein are necessarily assumed to be exposed to water. By design, the profile program does not differentiate between globular and membrane proteins. In consequence, the side chains of membrane proteins projecting outward from the transmembrane segments would interact with the lipid membrane milieu, and ought to be considered buried, while the current algorithm may treat them as exposed. This trend can be gathered from the third panel in Fig. 8d, showing the environment of *Rhodobacter capsulatus* porin. For a visual impression, we arbitrarily converted the six side-chain environment categories (B1, B2, B3, P1, P2, and E) into respective environmental "hydrophobicities" [1-fraction polar) x (area buried)] using average values from Fig. 4 of (Bowie et al., 1991, Science, 253:164-170). In principl , each consecutive residue in a transmembrane β-strand might be expected to show a clear alt rnation in environment with r spect to the prior one. Some limited alternation is det ct d for

the strands (panel 3, Fig. 7d), but only rarely going into the high hydrophobicity region that would be expected for the bilayer environment. We believe that perhaps that is why the global scores we obtain in Figs. 9a and 9b are lower than those obtained for globular proteins, and why the algorithm does not separate the protein scores as it otherwise might. Still, even with limitations, the algorithm is promising in that it does some discrimination consistent with expectations.

Functional possibilities for multimers. Anti-porters such as the H^+/K^+-ATPase, plus symports such as the $Na^+/K^+/2Cl^-$ transporters pose as questions whether the multiply transported ions might share the same route through the protein, and how could that be, especially for ions of opposite charge. Consideration of the porin arrangement leads us to speculate that perhaps the paths for the individual species might be separate, after all; each species might traverse the channel of a different "repeat", each one having its own suitable selectivity. Merging of the channels might account somehow for the stoichiometry observed.

Various publications are cited herein, the texts of which are hereby incorporated by reference in their entireties.

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT: Fischbarg, Jorge
Czegledy, Ferenc
Iserovich, Pavel
Li, Jun
Cheung, Min

(ii) TITLE OF INVENTION: A METHOD FOR PREDICTING PROTEIN STRUCTURE

(iii) NUMBER OF SEQUENCES: 3

(iv) CORRESPONDENCE ADDRESS:
(A) ADDRESSEE: Brumbaugh, Graves, Donohue & Raymond
(B) STREET: 30 Rockefeller Plaza
(C) CITY: New York
(D) STATE: NY
(E) COUNTRY: USA
(F) ZIP: 10112-0228

(v) COMPUTER READABLE FORM:
(A) MEDIUM TYPE: Floppy disk
(B) COMPUTER: IBM PC compatible
(C) OPERATING SYSTEM: PC-DOS/MS-DOS
(D) SOFTWARE: PatentIn Release #1.0, Version #1.25

(vi) CURRENT APPLICATION DATA:
(A) APPLICATION NUMBER: US 08/355,844
(B) FILING DATE: 14-DEC-1994
(C) CLASSIFICATION:

(viii) ATTORNEY/AGENT INFORMATION:
(A) NAME: Tang, Henry Y.S.
(B) REGISTRATION NUMBER: 29,705
(C) REFERENCE/DOCKET NUMBER: A29927-50/29910

(ix) TELECOMMUNICATION INFORMATION:
(A) TELEPHONE: 212-408-2586
(B) TELEFAX: 212-765-2519

(2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 340 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(vi) ORIGINAL SOURCE:
(A) ORGANISM: Escherichia coli

(ix) FEATURE:
(A) NAME/KEY: Peptide
(B) LOCATION: 1..340
(C) OTHER INFORMATION: OmpF porin protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

```
Ala Glu Ile Tyr Asn Lys Asp Gly Asn Lys Val Asp Leu Tyr Gly Lys
1               5                  10                  15

Ala Val Gly Leu His Tyr Arg Ser Lys Gly Asn Gly Glu Asn Ser Tyr
            20                  25                  30

Gly Gly Asn Gly Asp Met Thr Tyr Ala Arg Leu Gly Phe Lys Gly Glu
        35                  40                  45

Thr Gln Ile Asn Ser Asp Leu Thr Gly Tyr Gly Gln Trp Glu Tyr Asn
    50                  55                  60

Phe Gln Gly Asn Asn Ser Glu Gly Ala Asp Ala Gln Thr Gly Asn Lys
65                  70                  75                  80

Thr Arg Leu Ala Phe Ala Gly Leu Lys Tyr Ala Asp Val Gly Ser Phe
                85                  90                  95

Asp Tyr Gly Arg Asn Tyr Gly Val Val Tyr Asp Ala Leu Gly Tyr Thr
            100                 105                 110

Asp Met Leu Pro Glu Phe Gly Gly Asp Thr Ala Tyr Ser Asp Asp Phe
        115                 120                 125

Phe Val Gly Arg Val Gly Gly Val Ala Thr Tyr Arg Asn Ser Asn Phe
    130                 135                 140

Phe Gly Leu Val Asp Gly Leu Asn Phe Ala Val Gln Tyr Leu Gly Lys
145                 150                 155                 160

Asn Glu Arg Asp Thr Ala Arg Arg Ser Asn Gly Asp Gly Val Gly Gly
                165                 170                 175

Ser Ile Ser Tyr Glu Tyr Asx Gly Phe Gly Ile Val Gly Ala Tyr Gly
            180                 185                 190

Ala Ala Asp Arg Thr Asn Leu Gln Glu Ala Gln Pro Leu Gly Asn Gly
        195                 200                 205

Lys Lys Ala Glu Gln Trp Ala Thr Gly Leu Lys Tyr Asp Ala Asn Asn
    210                 215                 220

Ile Tyr Leu Ala Ala Asn Tyr Gly Glu Thr Arg Asn Ala Thr Pro Ile
225                 230                 235                 240

Thr Asn Lys Phe Thr Asn Thr Ser Gly Phe Ala Asn Lys Thr Gln Asp
                245                 250                 255

Val Leu Leu Val Ala Gln Tyr Gln Phe Asp Phe Gly Leu Arg Pro Ser
            260                 265                 270

Ile Ala Tyr Thr Lys Ser Lys Ala Lys Asp Val Glu Gly Ile Gly Asp
        275                 280                 285

Val Asp Leu Val Asn Tyr Ph  Glu Val Gly Ala Thr Tyr Tyr Phe Asn
    290                 295                 300

Lys Asn Met Ser Thr Tyr Val Asp Tyr Ile Ile Asn Gln Ile Asp Ser
305                 310                 315                 320

Asp Asn Lys Leu Gly Val Gly Ser Asp Asp Thr Val Ala Val Gly Il
                325                 330                 335
```

Thr Ala Val Asp His Lys Ala Tyr Gly Leu S r Val Asp Ser Thr Phe
225 230 235 240

Gly Ala Thr Thr Val Gly Gly Tyr Val Gln Val Leu Asp Ile Asp Thr
245 250 255

Ile Asp Asp Val Thr Tyr Tyr Gly Leu Gly Ala Ser Tyr Asp Leu Gly
260 265 270

Gly Gly Ala Ser Ile Val Gly Gly Ile Ala Asp Asn Asp Leu Pro Asn
275 280 285

Ser Asp Asn Val Ala Asp Leu Gly Val Lys Phe Lys Phe
290 295 300

(2) INFORMATION FOR SEQ ID NO:3:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 492 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(vi) ORIGINAL SOURCE:
(A) ORGANISM: Human

(ix) FEATURE:
(A) NAME/KEY: Peptide
(B) LOCATION: 1..492
(C) OTHER INFORMATION: Facilitative glucose transportor Glut1 protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:

Met Glu Pro Ser Ser Lys Lys Leu Thr Gly Arg Leu Met Leu Ala Val
1 5 10 15

Gly Gly Ala Val Leu Gly Ser Leu Gln Phe Gly Tyr Asn Thr Gly Val
20 25 30

Ile Asn Ala Pro Gln Lys Val Ile Glu Glu Phe Tyr Asn Gln Thr Trp
35 40 45

Val His Arg Tyr Gly Glu Ser Ile Leu Pro Thr Thr Leu Thr Thr Leu
50 55 60

Trp Ser Leu Ser Val Ala Ile Phe Ser Val Gly Gly Met Ile Gly Ser
65 70 75 80

Phe Ser Val Gly Leu Phe Val Asn Arg Phe Gly Arg Arg Asn Ser Met
85 90 95

Leu Met Met Asn Leu Leu Ala Phe Val Ser Ala Val Leu Met Gly Phe
100 105 110

Ser Lys Leu Gly Lys Ser Phe Glu Met Leu Ile Leu Gly Arg Phe Ile
115 120 125

Ile Gly Val Tyr Cys Gly Leu Thr Thr Gly Phe Val Pro Met Tyr Val
130 135 140

Gly Glu Val Ser Pro Thr Ala Phe Arg Gly Ala L u Gly Thr Leu His
145 150 155 160

Gln Leu Gly Ile Val Val Gly Ile Leu Ile Ala Gln Val Phe Gly Leu
165 170 175

Asp Ser Ile Met Gly Asn Lys Asp Leu Trp Pro Leu Leu Leu Ser Ile
180 185 190

Ile Phe Ile Pro Ala Leu Leu Gln Cys Ile Val Ile Pro Phe Cys Pro
195 200 205

Glu Ser Pro Arg Phe Leu Leu Ile Asn Arg Asn Glu Glu Asn Arg Ala
210 215 220

Lys Ser Val Leu Lys Lys Leu Arg Gly Thr Ala Asp Val Thr His Asp
225 230 235 240

Leu Gln Glu Met Lys Glu Glu Ser Arg Gln Met Met Arg Glu Lys Lys
245 250 255

Val Thr Ile Leu Glu Leu Phe Arg Ser Pro Ala Tyr Arg Gln Pro Ile
260 265 270

Leu Ile Ala Val Val Leu Gln Leu Ser Gln Gln Leu Ser Gly Ile Asn
275 280 285

Ala Val Phe Tyr Tyr Ser Thr Ser Ile Phe Glu Lys Ala Gly Val Gln
290 295 300

Gln Pro Val Tyr Ala Thr Ile Gly Ser Gly Ile Val Asn Thr Ala Phe
305 310 315 320

Thr Val Val Ser Leu Phe Val Val Glu Arg Ala Gly Arg Arg Thr Leu
325 330 335

His Leu Ile Gly Leu Ala Gly Met Ala Gly Gln Ala Ile Leu Met Thr
340 345 350

Ile Ala Leu Ala Leu Leu Glu Gln Leu Pro Trp Met Ser Tyr Leu Ser
355 360 365

Ile Val Ala Ile Phe Gly Phe Val Ala Phe Phe Glu Val Gly Pro Gly
370 375 380

Pro Ile Pro Trp Phe Ile Val Ala Glu Leu Glu Ser Gln Gly Pro Arg
385 390 395 400

Pro Ala Ala Ile Ala Val Ala Gly Phe Ser Asn Trp Thr Ser Asn Phe
405 410 415

Ile Val Gly Met Cys Phe Gln Tyr Val Glu Gln Leu Cys Gly Pro Tyr
420 425 430

Val Phe Ile Ile Phe Thr Val Leu Leu Val Leu Phe Phe Ile Arg Thr
435 440 445

Tyr Phe Lys Val Pro Glu Thr Lys Gly Arg Thr Phe Asp Glu Ile Ala
450 455 460

Ser Gly Phe Arg Gln Gly Gly Ala Ser Gln Ser Asp Lys Thr Pr Glu
465 470 475 480

Glu Leu Phe His Pro Leu Gly Ala Asp Ser Gln Val
485 490

Claims

1. A method of predicting the tendency of a protein to form an amphiphilic α structure,

comprising calculating a series of values for $U_{\alpha x}$ for a series of portions of the protein, each portion having a span of x residues, wherein the series of portions spans the protein, and wherein x is any integer, comprising calculating a value for $U_{\alpha x}$ using the equation $U_{\alpha x} = H_x + \mu_{\alpha x}$ - <pt>, wherein H_x is the average hydrophobicity for a span of x residues using the Kyte-Doolitte scale, $\mu_{\alpha x}$ is the hydrophobic moment (span x) for α structures, the angle between one residue and the successive residue being 100°, and <pt> is the position dependent turn propensity, and further comprising depicting the values for $U_{\alpha x}$ graphically to form a series of peaks, wherein peaks wide enough to correspond to a segment of the amino acid sequence long enough to span the membrane as an α-helix are predicted to be α structures.

2. The method according to claim 1, using the source code set forth in pages 12 - 18 of the specification.

3. The method according to claim 1, using the source code set forth in pages 20 - 32 of the specification.

4. The method according to claim 1, where x has a value of seven.

5. Th method according to claim 2, where x has a value of seven.

6. The method according to claim 3, where x has a value of seven.

7. The method according to claim 1, where x has a value of twenty-one.

8. The method according to claim 2, where x has a value of twenty-one.

9. The method according to claim 3, where x has a value of twenty-one.

10. A method of predicting the tendency of a protein to form an amphiphilic β structure,

comprising calculating a series of values for $U_{\beta x}$ for a series of portions of the protein, each portion having a span of x residues, wherein the series of portions spans the protein, and wherein x is any integer, comprising calculating a value for $U_{\beta x}$ using the equation $U_{\beta x} = H_x + \mu_{\beta x} - <pt>$, wherein H_x is the average hydrophobicity for a span of x residues using the Kyte-Doolitte scale, $\mu_{\beta x}$ is the hydrophobic moment (span x) for β structures, the angle between one residue and the successive residue being 160°, and <pt> is the position dependent turn propensity, and further comprising depicting the values for $U_{\beta x}$ graphically to form a series of peaks, wherein peaks that are too narrow to correspond to a segment of the amino acid sequence long enough to span the membran as an α-helix but which are wid enough to correspond to a s gment of the amino acid sequence with a length between 6 and 14 amino acid residues are predicted to be β structures.

11. The m thod according to claim 10, using the algorithm set forth in pages 12 - 18 of the specification.

12. The method according to claim 10, using the algorithm set forth in pages 20 - 32 of the specification.

13. The method according to claim 10, where *x* has a value of seven.

14. The method according to claim 11, where *x* has a value of seven.

15. The method according to claim 12, where *x* has a value of seven.

16. The method according to claim 10, where *x* has a value of twenty-one.

17. The method according to claim 11, where *x* has a value of twenty-one.

18. The method according to claim 12, where *x* has a value of twenty-one.

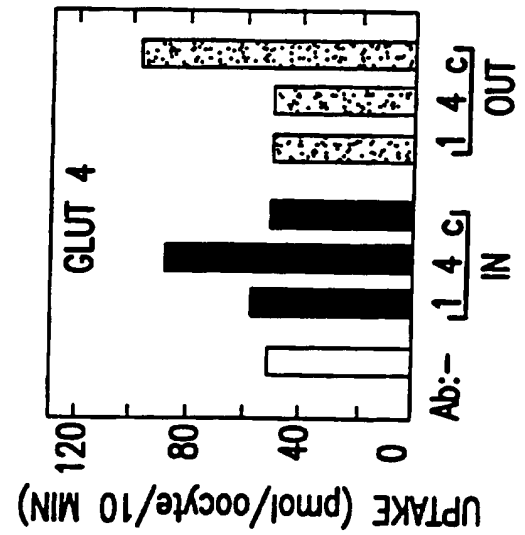


FIG.1c

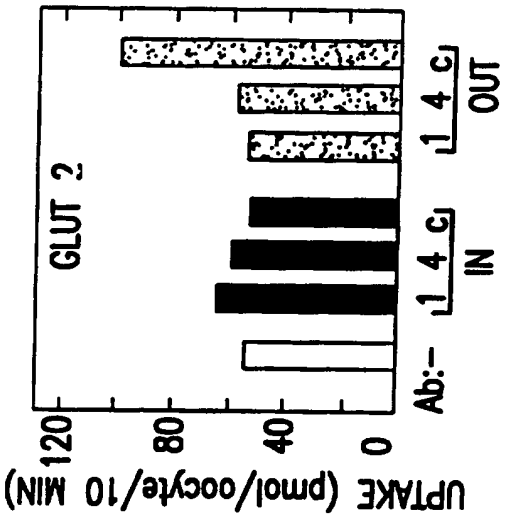


FIG.1b

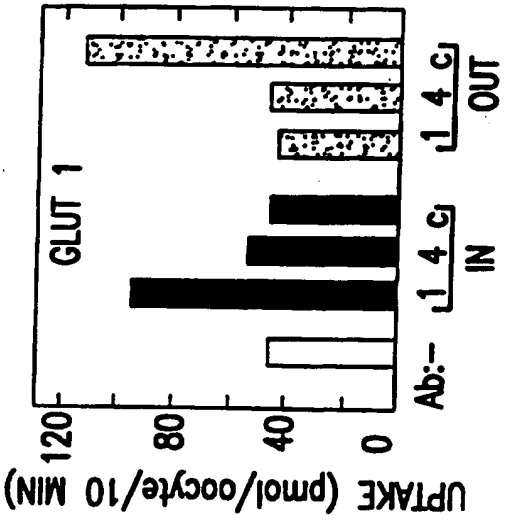


FIG.1a

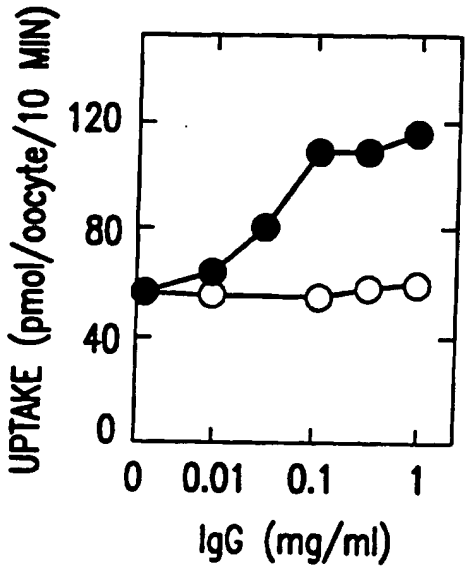


FIG.1d

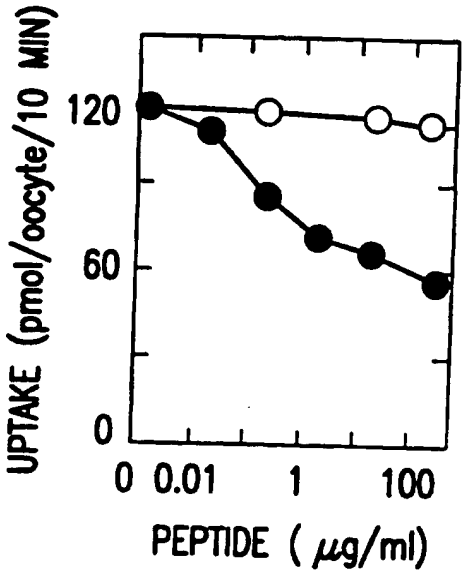


FIG.1e

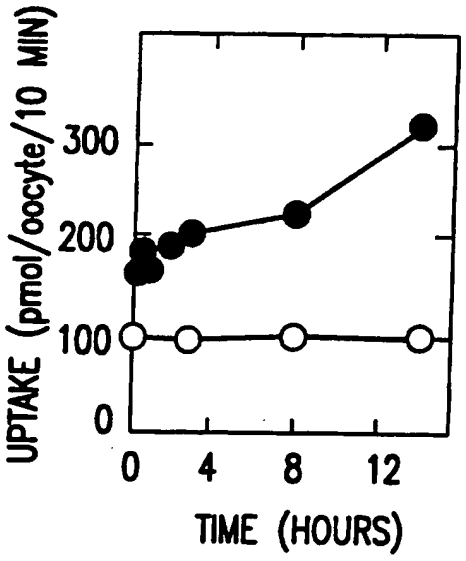


FIG.1f

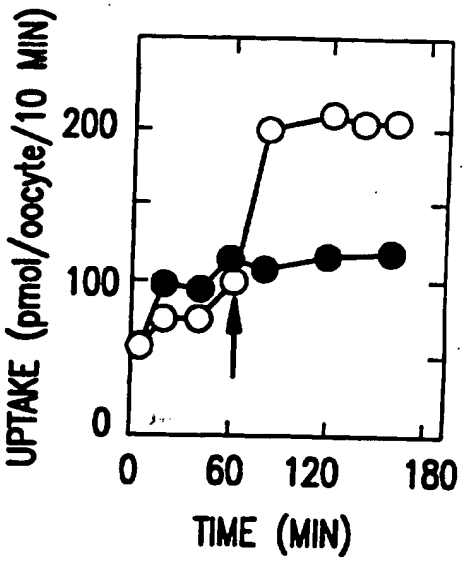


FIG.1g

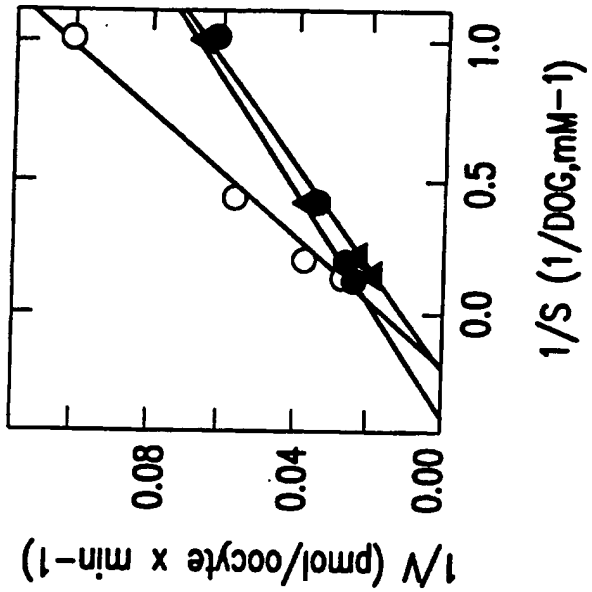

1/V (pmol/oocyte x min-1)
0.08
0.04
0.00
0.0
0.5
1.0
1/S (1/DOG,mM-1)

FIG. 1i

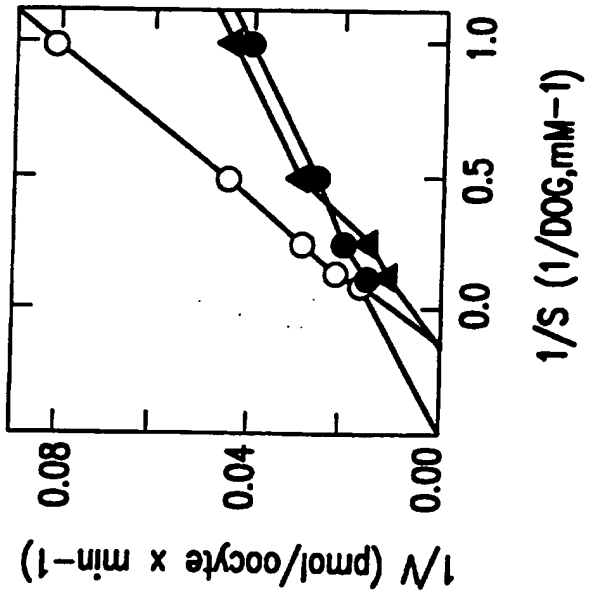

1/V (pmol/oocyte x min-1)
0.08
0.04
0.00
0.0
0.5
1.0
1/S (1/DOG,mM-1)

FIG. 1h

```
Ompf        ..... ..... ..... ..... .....
S16070      ..... ..... ..... ..... .....
Gtr1_Human  MEPSSKKLTG RLMLAVGGAV LGSLQFGYNT GVINAPQKVI EEFYNQTVVH 50

Ompf        ....DGNKVD LYGKA.VGLH YESKGNGENS YG.....GNG DMTYARLGFK
S16070      ......EVK LSGDARMGVM Y...NGDDWN FS.....SRS .....RVLFT
Gtr1_Human  RYGESILPTT LTTLWSLSVA IRSVGGMIGS FSVGLFVNRF GRRNSMLMMN 100

Ompf        GETQINSDLT GYGQWEYNFQ GN.....NS EGADAQTGNK TRLAFAGLKY
S16070      MSGTTDSGL. ...EFGASFK AH.....ES VGARTGEDGT VFLSGAFGKI
Gtr1_Human  LLAFVSAVLM GFSKLGKSFE MLILGRFIIG VYCGLTTGFV PMYVGEVSPT 150

Ompf        ADVGSFDYGR N....YGVVY DALGYWDMLP EFGGDTAY.. SDDFFVGRVG
S16070      EMGDAIGASE AL...FGDLY E.VGYTDLDD RGGNDIPYLT GDERLTAEDN
Gtr1_Human  AFRGALGTLH QLGIVVGILI AQVFGLDSIM GNKDLWPLLL SIIFIPALLQ 200

Ompf        GVAT..YRNS NFFGLV..DG LNFAVQYLGK NERDTARRSN GDGVGGSISY
S16070      PVLL..YTYS .....A..GA FSVAAS.MSD GKVGETSEDD AQEMAVAAAY
Gtr1_Human  CIVDPFCPES PRFLLINRNE ENRAKSVLKK LRGTADVTHD LQEMKEESRQ 250

Ompf        EYEG..FGIV GAYGAADRTN LQEAQPLGNG KKAEQWATGL KYDANNIYLA
S16070      TFGN..YTVG LGYEKIDSPD ....TALMAD MEQLELAAIA KFGATNV..K
Gtr1_Human  MMREKKVTIL ELERSPAYRQ PILIAVVLQL SQQLSGINAV FYYSTSIFEK 300

Ompf        ANYGETRNAT PITNKFTNTS GFANKTQDVL LVAQYQFDFG LRPSI.AYTK
S16070      AYYADGELDR DFARAVFDLT PVAAAATAVD HKA...YGLS VDSTFGATTV
Gtr1_Human  AGVQQPVYAT .IGSGIVNTA PTVVSLFVVE RAGRRTLHLI GLAGMAGGAI 350

Ompf        SKAKDVEGIG DVDLVNYFEV GATY....YF NKNMSTYVDY IINQIDSDNK
S16070      GGYVQVLDID TIDDVTYYGL GASY....DL GGGAS..... IVGGI.ADND
Gtr1_Human  LMTIALALLE QLPWMSYLSI VAIFGFVAFF EVGPGPIPWF IVAELFSQGP 400

Ompf        LGVGSDDTV. .......AVG IVYQFAEIYN K..... .....
S16070      LP.NSDNVA. .......DLG VKFKF..... ..... .....
Gtr1_Human  RPAAIAVAGF SNWTSNFIVG MCFQYVEQLC GPYVFIIFTV LLVLFFIRTY 450

Ompf        ..... ..... ..... ..... ...
S16070      ..... ..... ..... ..... ...
Gtr1_Human  FKVPETKGFT FDEIASGFRQ GGASQSDKTP EELFHPLGAD SQV        493
```

FIG.2

FIG.3

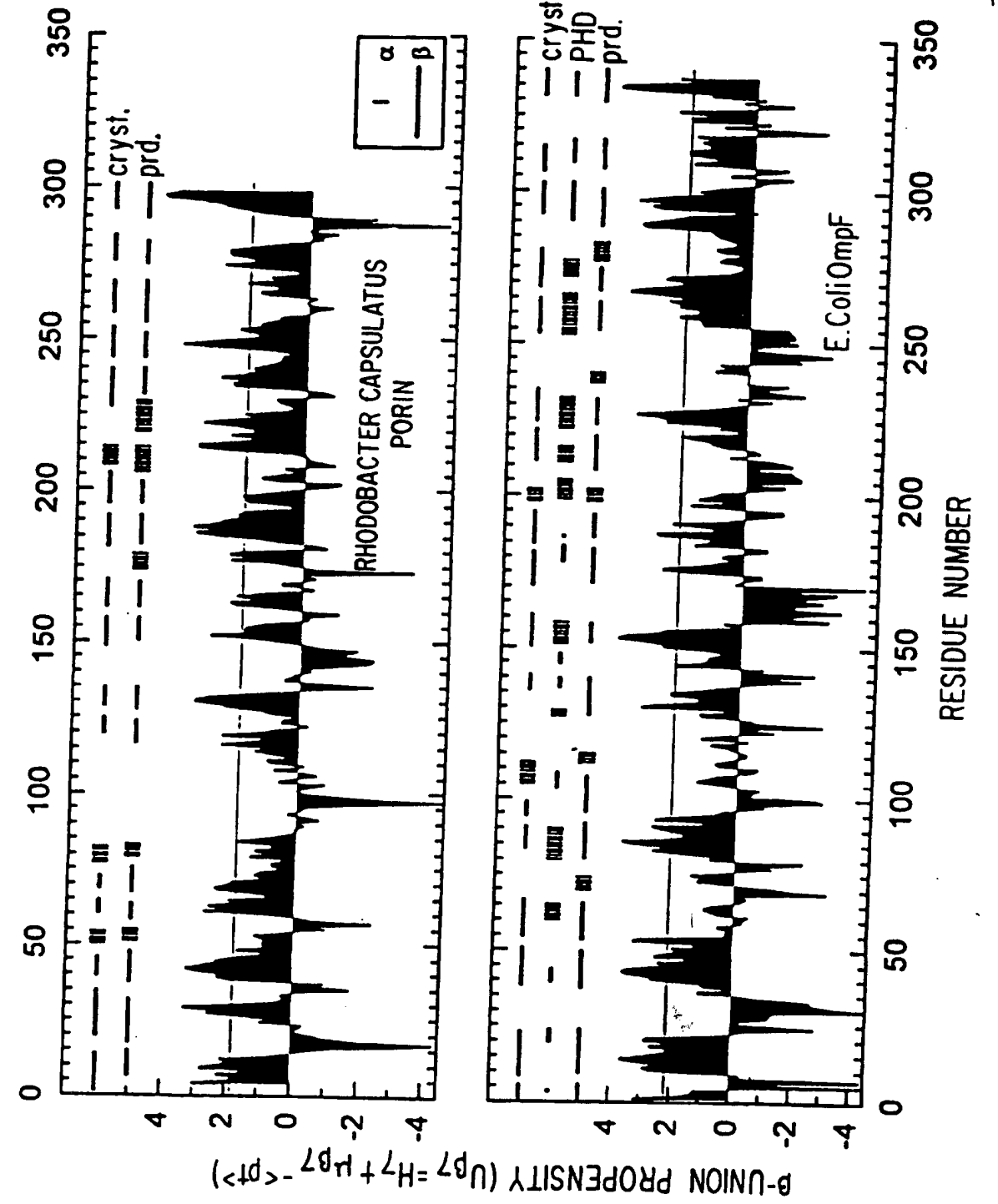
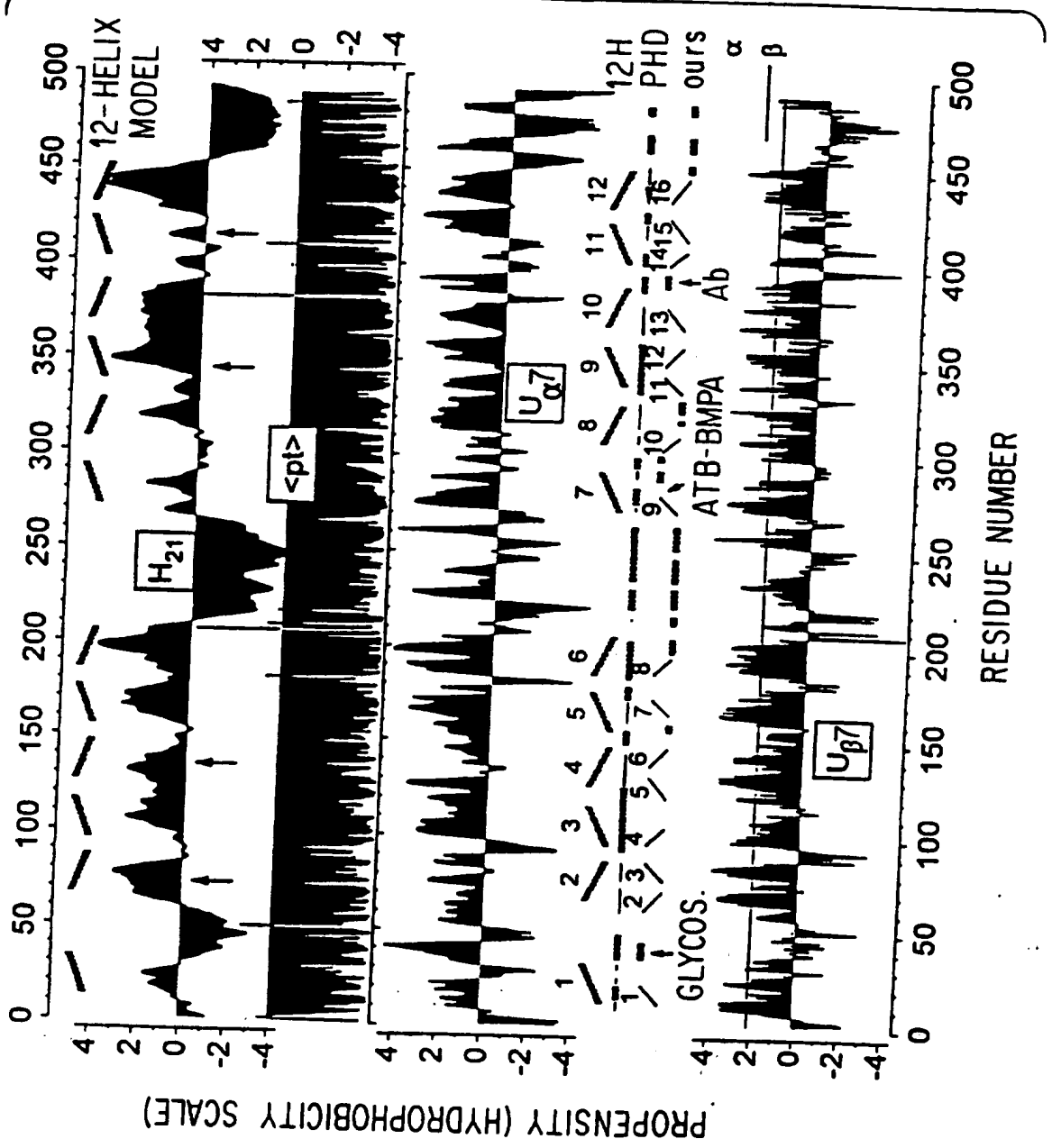

RHODOBACTER CAPSULATUS
PORIN
E.Coli OmpF
cryst.
prd.
PHD
α
β
RESIDUE NUMBER
β-UNION PROPENSITY

FIG. 4


12-HELIX MODEL
H21
<pt>
Uα7
Uβ7
12H PHD
ours
ATB-BMPA
Ab
GLYCOS.
α
β
RESIDUE NUMBER
PROPENSITY (HYDROPHOBICITY SCALE)

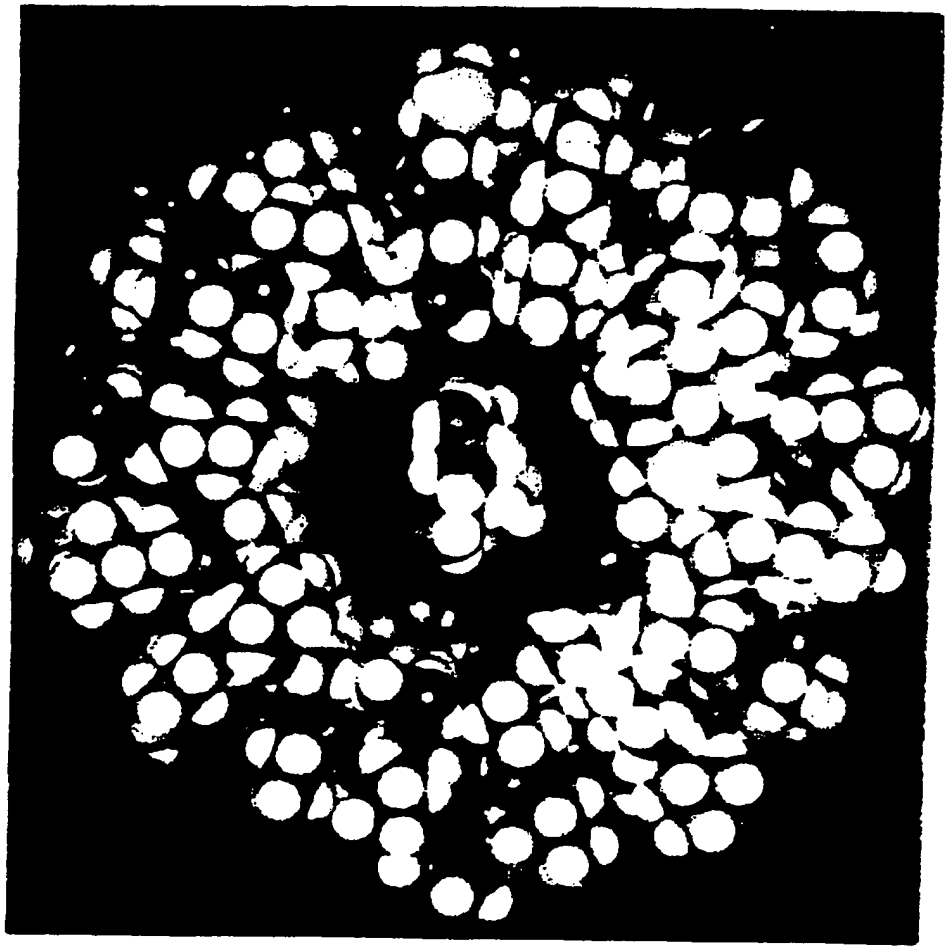


FIG.5

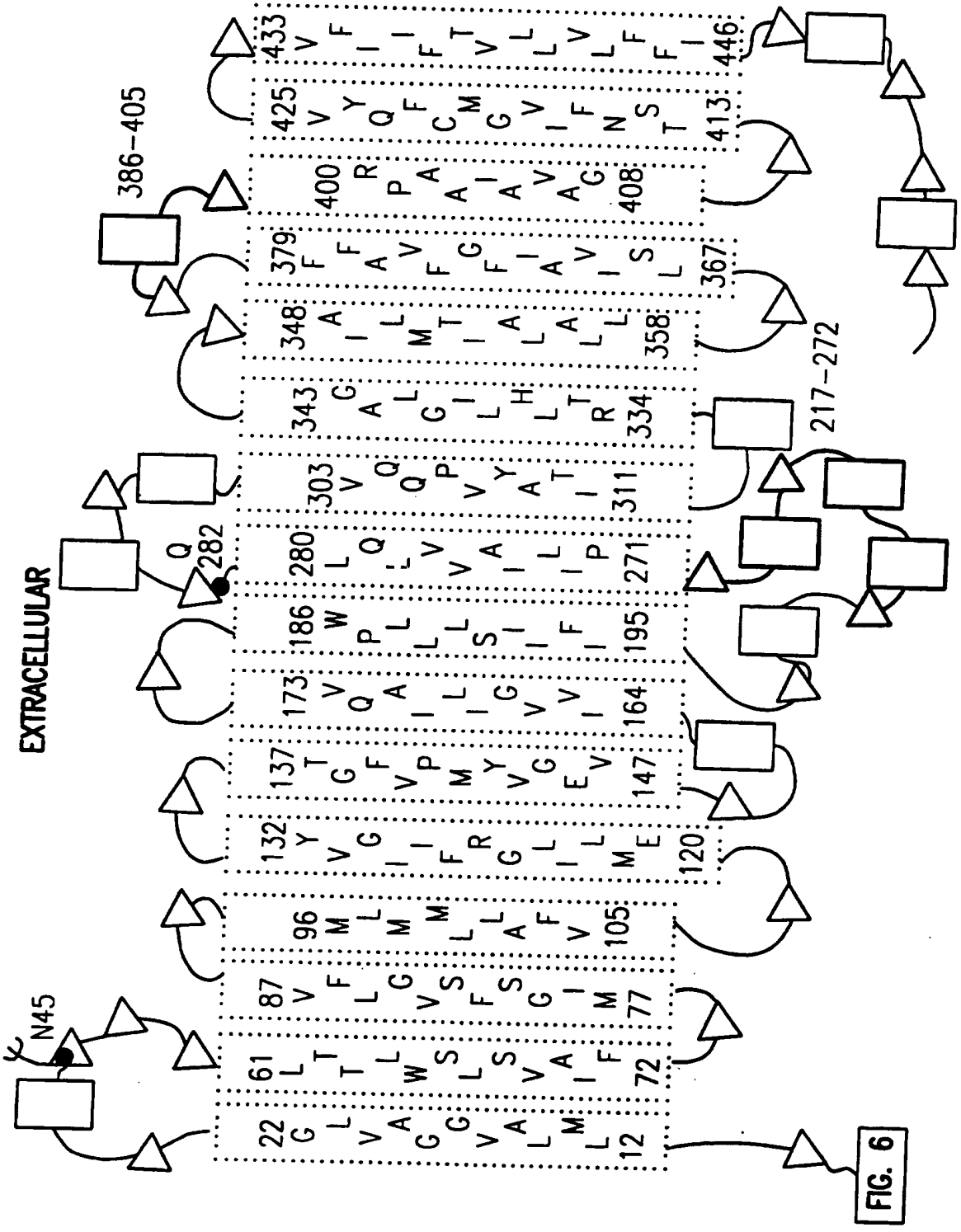

EXTRACELLULAR
N45
Q 282
386–405
217–272
22 G L V A G G V A L M L 12
61 L T T L W S L S V A I F 72
87 V F L G V S F S G I M 77
96 M L M M L L A F V 105
132 Y V G I I F R G L I L M E 120
137 T G F V P M Y V G V E V 147
173 V V Q A I L L G V V I 164
186 W P L L L S I I F I 195
280 L Q L V V A I L I P 271
303 V Q Q P V Y A T I 311
343 G A L G I L H L T R 334
348 A I L M T I A L A L L 358
379 F F A V F G F I A V I S 367
400 R P A A I A V A G 408
425 V Y Q F C M G V I F N S T 413
433 V F I I F T V L V L F I 446

FIG. 6

FIG.6

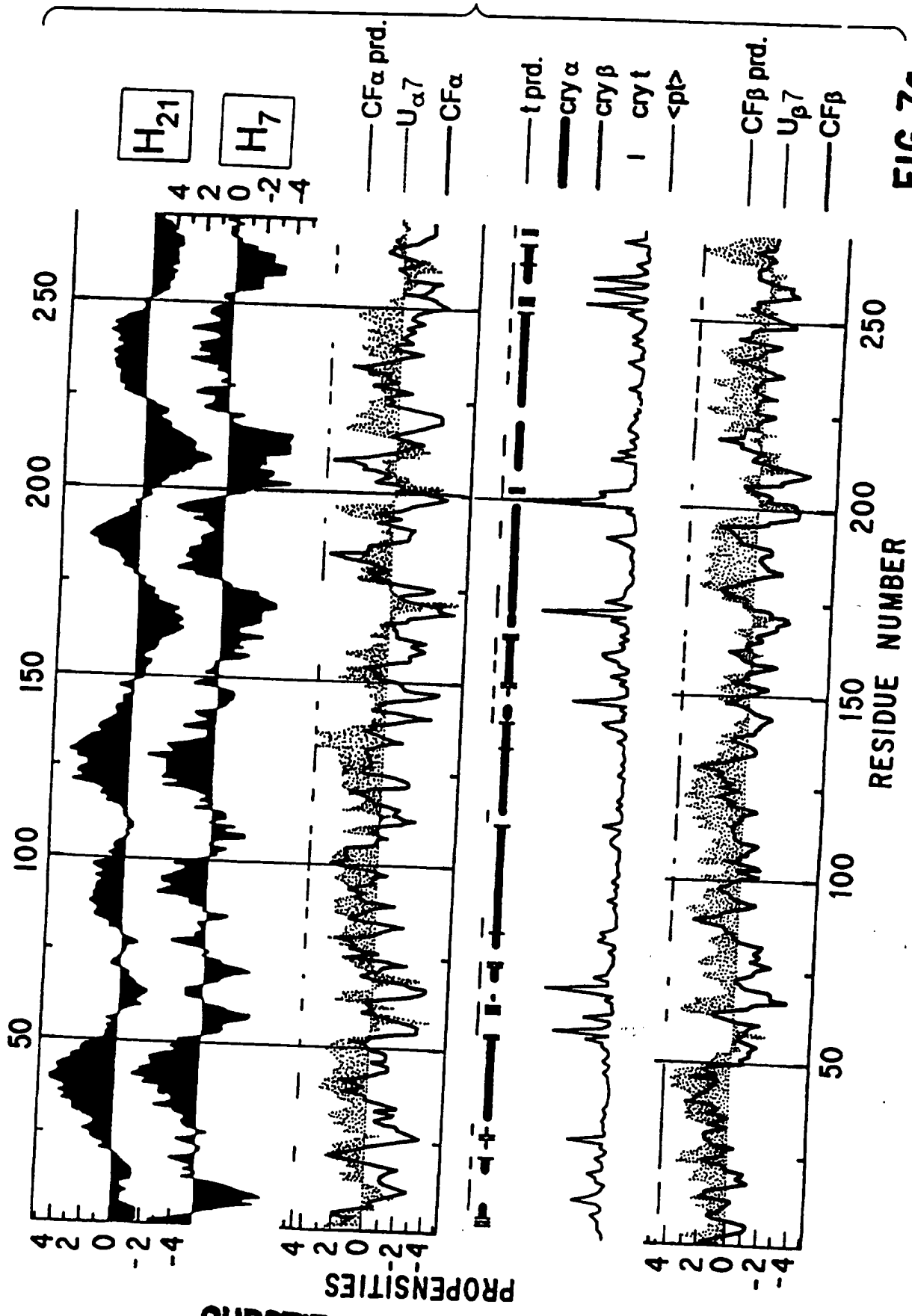

H21
H7
CFα prd.
Uα7
CFα
t prd.
cryα
cryβ
cryt
<pt>
CFβ prd.
Uβ7
CFβ
PROPENSITIES
RESIDUE NUMBER

FIG. 7a

FIG.7b

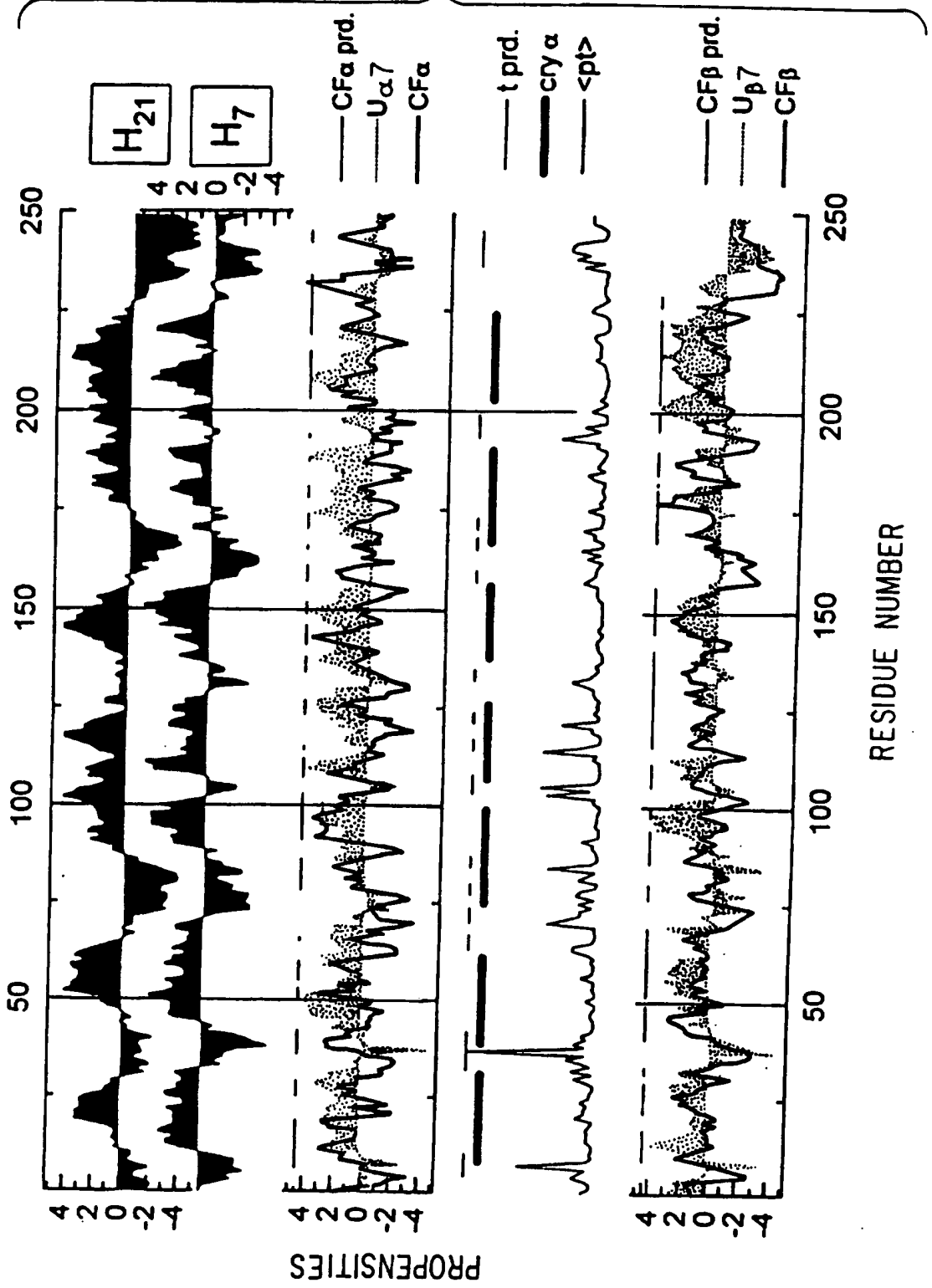
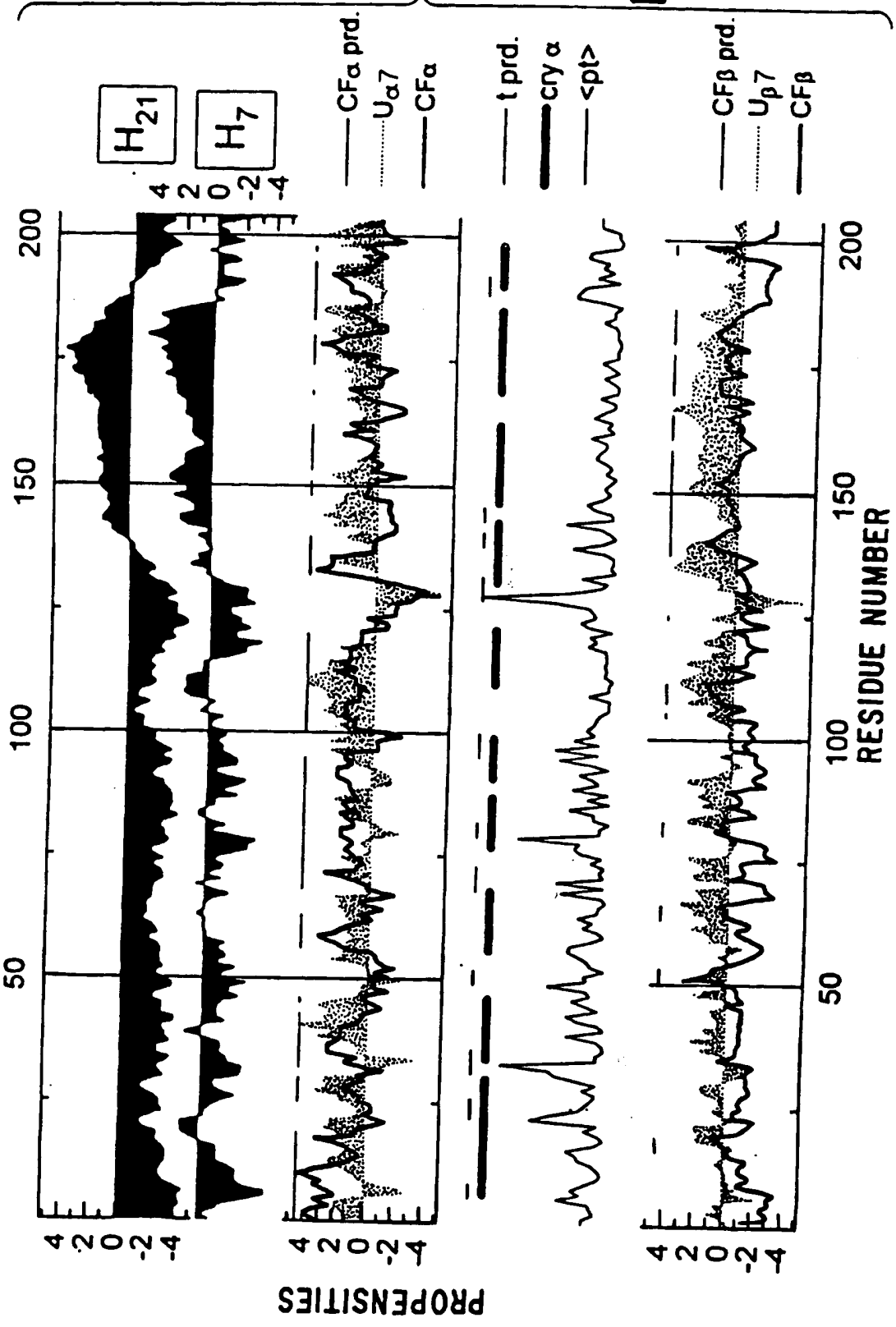

H21
H7
CFα prd.
Uα7
CFα
t prd.
cry α
<pt>
CFβ prd.
Uβ7
CFβ
RESIDUE NUMBER
PROPENSITIES

FIG. 7c


H21
H7
CFα prd.
Uα7
CFα
t prd.
cry α
<pt>
CFβ prd.
Uβ7
CFβ
4 2 0 -2 -4
200
150
100
50
RESIDUE NUMBER
PROPENSITIES

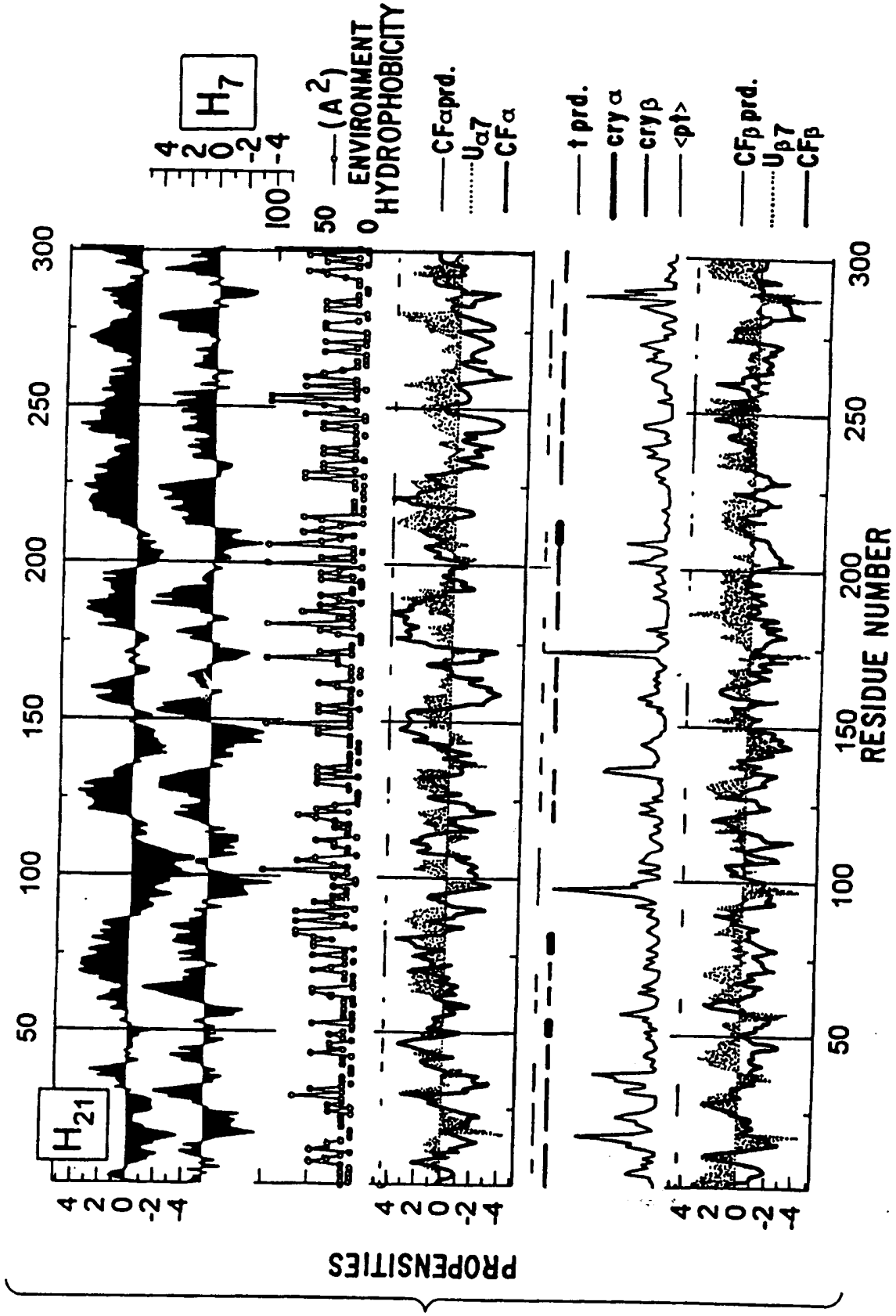

PROPENSITIES
RESIDUE NUMBER
ENVIRONMENT HYDROPHOBICITY
(A^2)
H_{21}
H_7
CFαprd.
$U_{\alpha 7}$
CFα
t prd.
cryα
cryβ
<pt>
CFβ prd.
$U_{\beta 7}$
CFβ

FIG. 7d

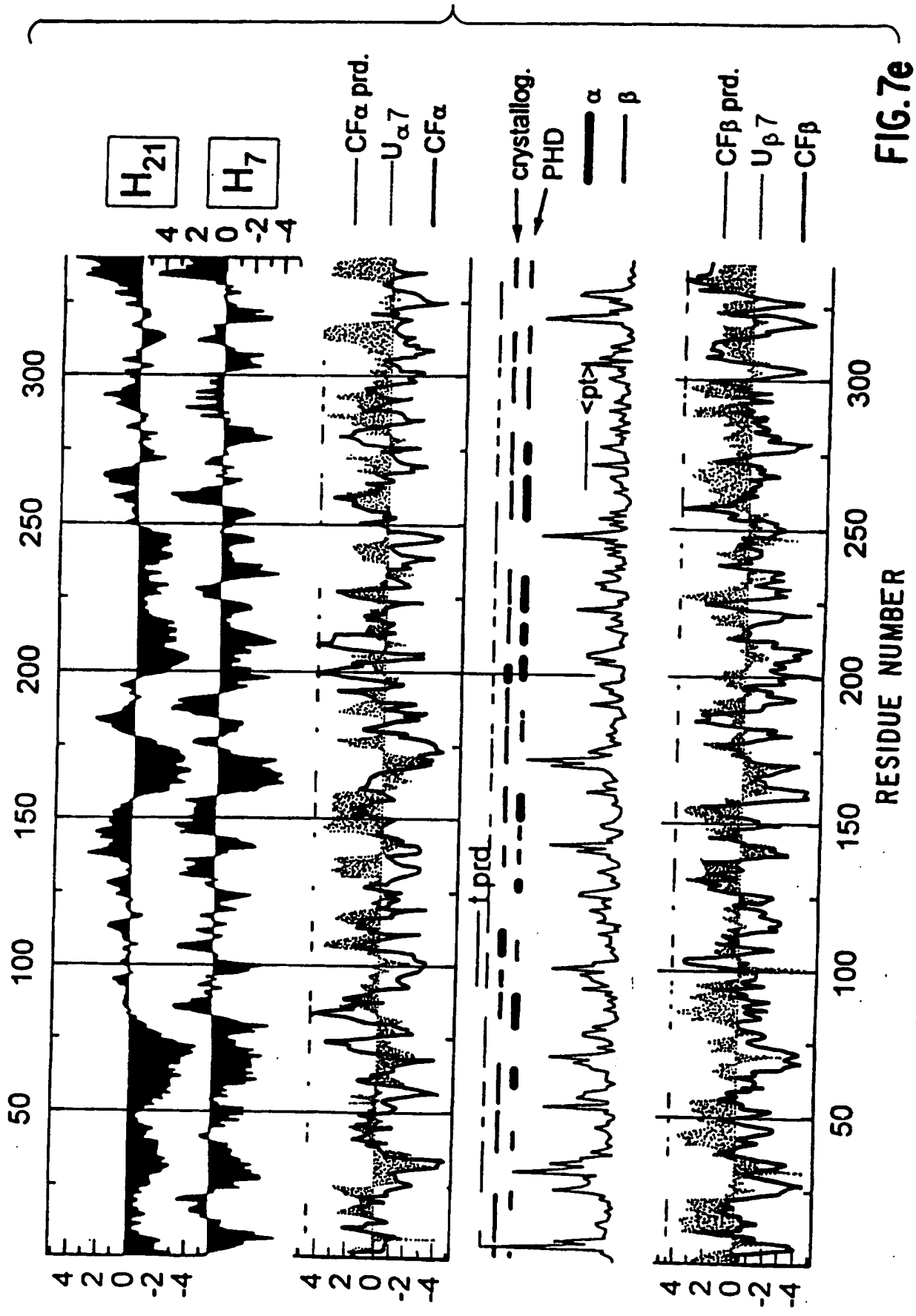

H21
H7
4 2 0 -2 -4
CFα prd.
Uα7
CFα
crystallog.
PHD
<pt>
α
β
CFβ prd.
Uβ7
CFβ
tprd
50 100 150 200 250 300
RESIDUE NUMBER
PROPENSITIES

FIG. 7e

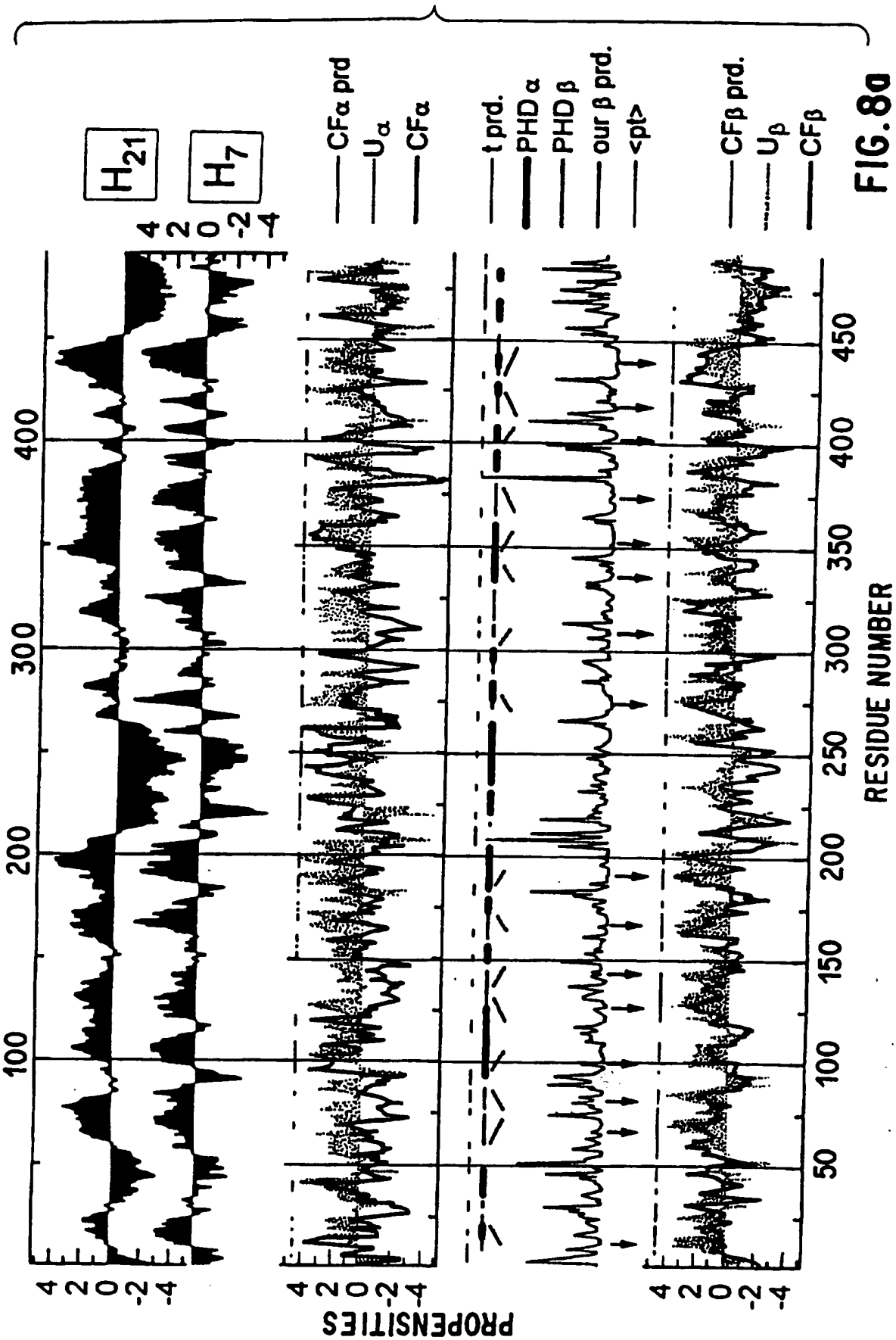

H21
H7
4 2 0 -2 -4
CFα prd
Uα
CFα
t prd.
PHD α
PHD β
our β prd.
<pt>
CFβ prd.
Uβ
CFβ
100
200
300
400
50
150
250
350
450
RESIDUE NUMBER
PROPENSITIES

FIG. 8a

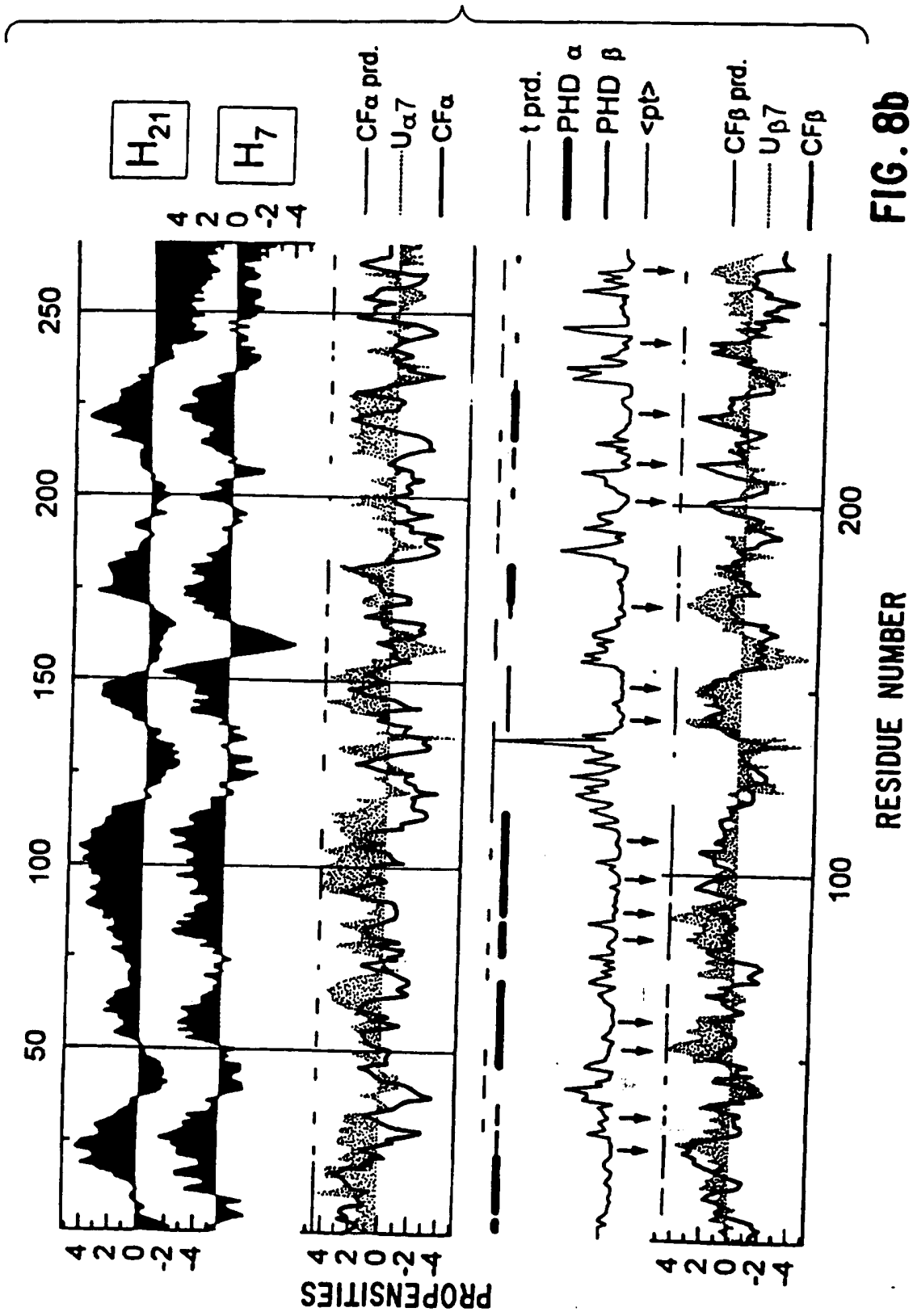
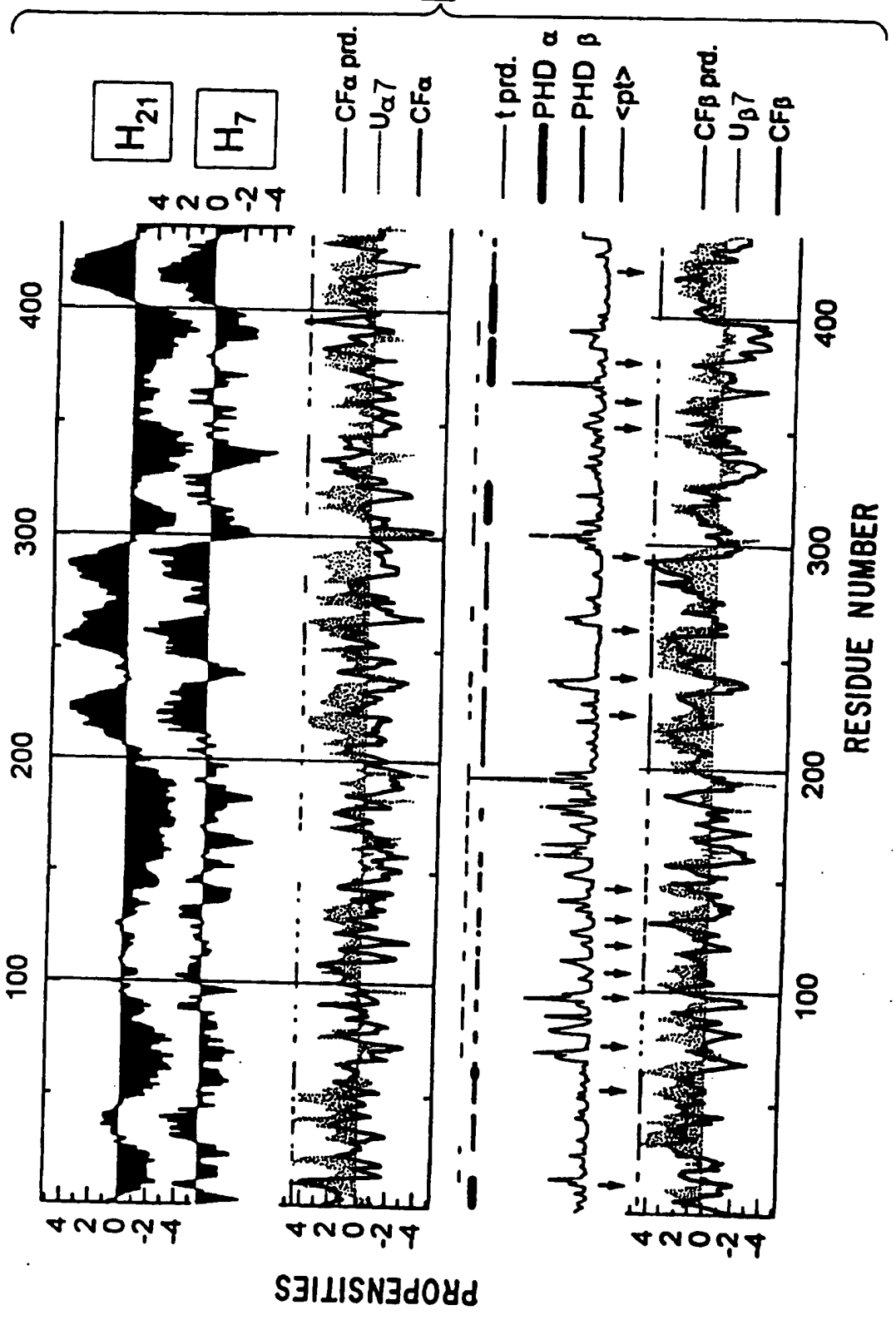

H21
H7
CFα prd.
Uα7
CFα
t prd.
PHD α
PHD β
<pt>
CFβ prd.
Uβ7
CFβ
PROPENSITIES
RESIDUE NUMBER

FIG. 8b

FIG.8c


H21
H7
CFα prd.
Uα7
CFα
t prd.
PHD α
PHD β
<pt>
CFβ prd.
Uβ7
CFβ
4 2 0 -2 -4
100
200
300
400
RESIDUE NUMBER
PROPENSITIES

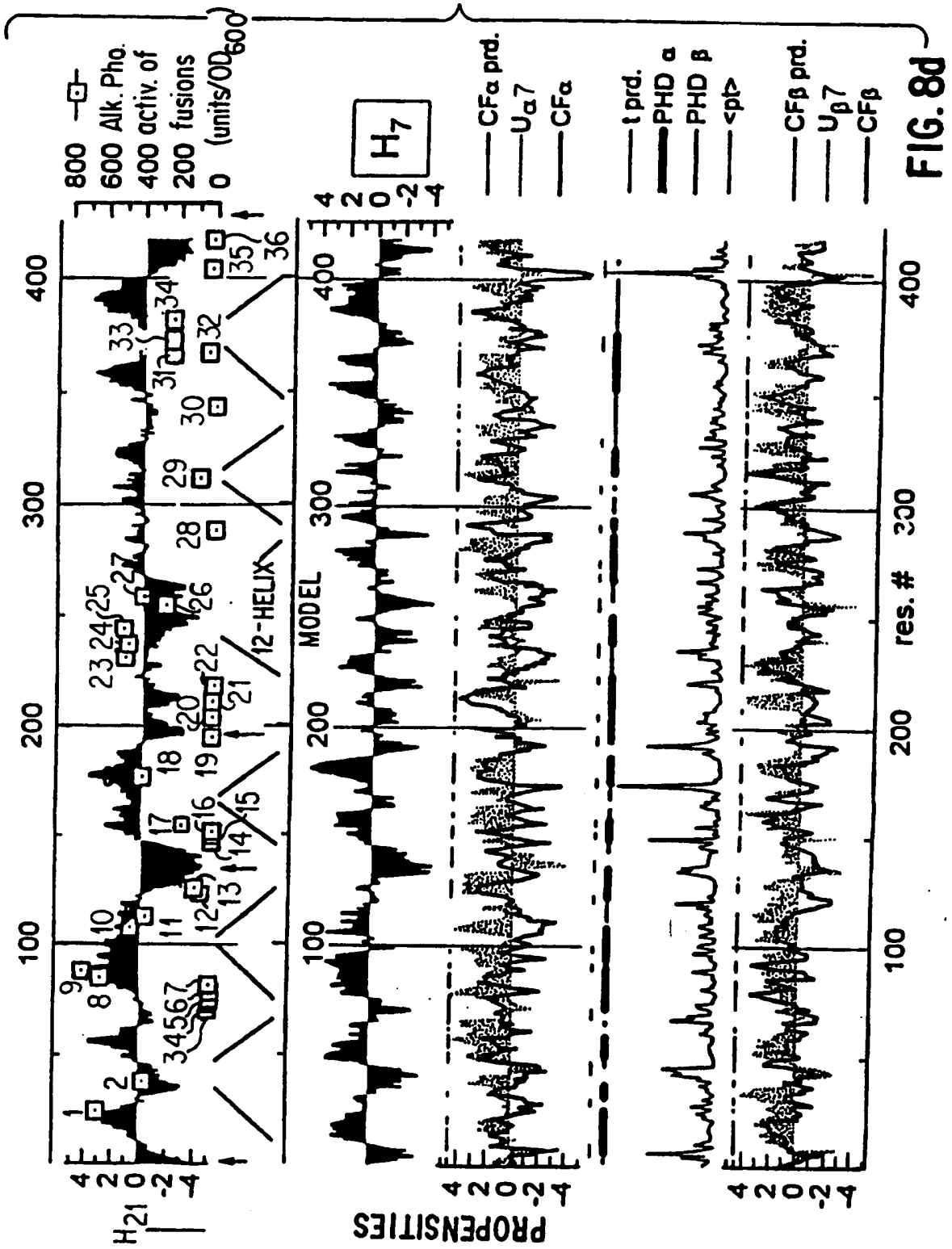

800
600 Alk. Pho.
400 activ. of
200 fusions
0 (units/OD)600
H7
CFα prd.
Uα7
CFα
t prd.
PHD α
PHD β
<pt>
CFβ prd.
Uβ7
CFβ
12-HELIX
MODEL
res. #
PROPENSITIES
H21

FIG. 8d

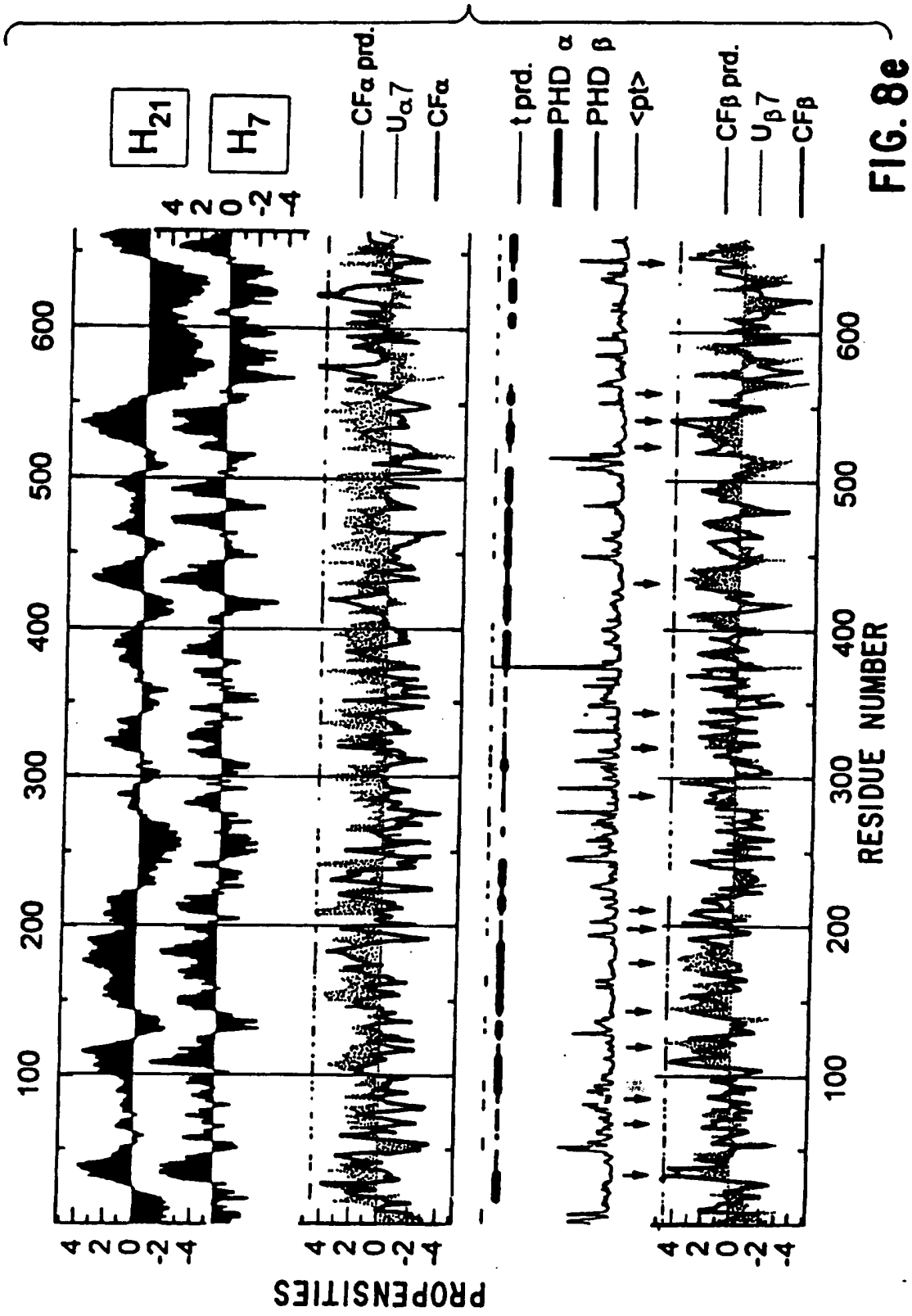

H21
H7
CFα prd.
Uα7
CFα
t prd.
PHD α
PHD β
<pt>
CFβ prd.
Uβ7
CFβ
4 2 0 -2 -4
100 200 300 400 500 600
RESIDUE NUMBER
PROPENSITIES

FIG. 8e

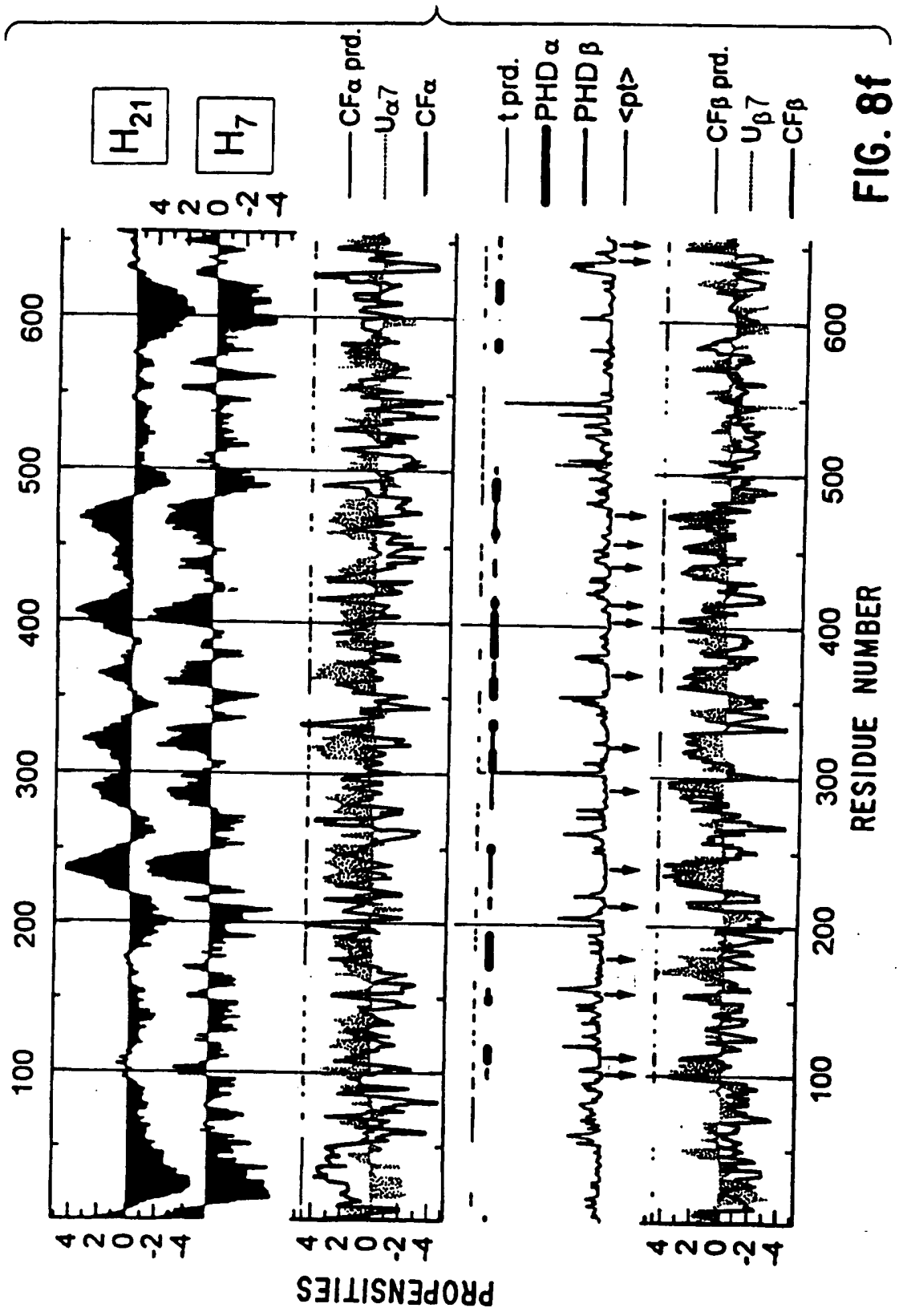

H21
H7
CFα prd.
Uα7
CFα
t prd.
PHD α
PHD β
<pt>
CFβ prd.
Uβ7
CFβ
4 2 0 -2 -4
100 200 300 400 500 600
RESIDUE NUMBER
PROPENSITIES

FIG. 8f

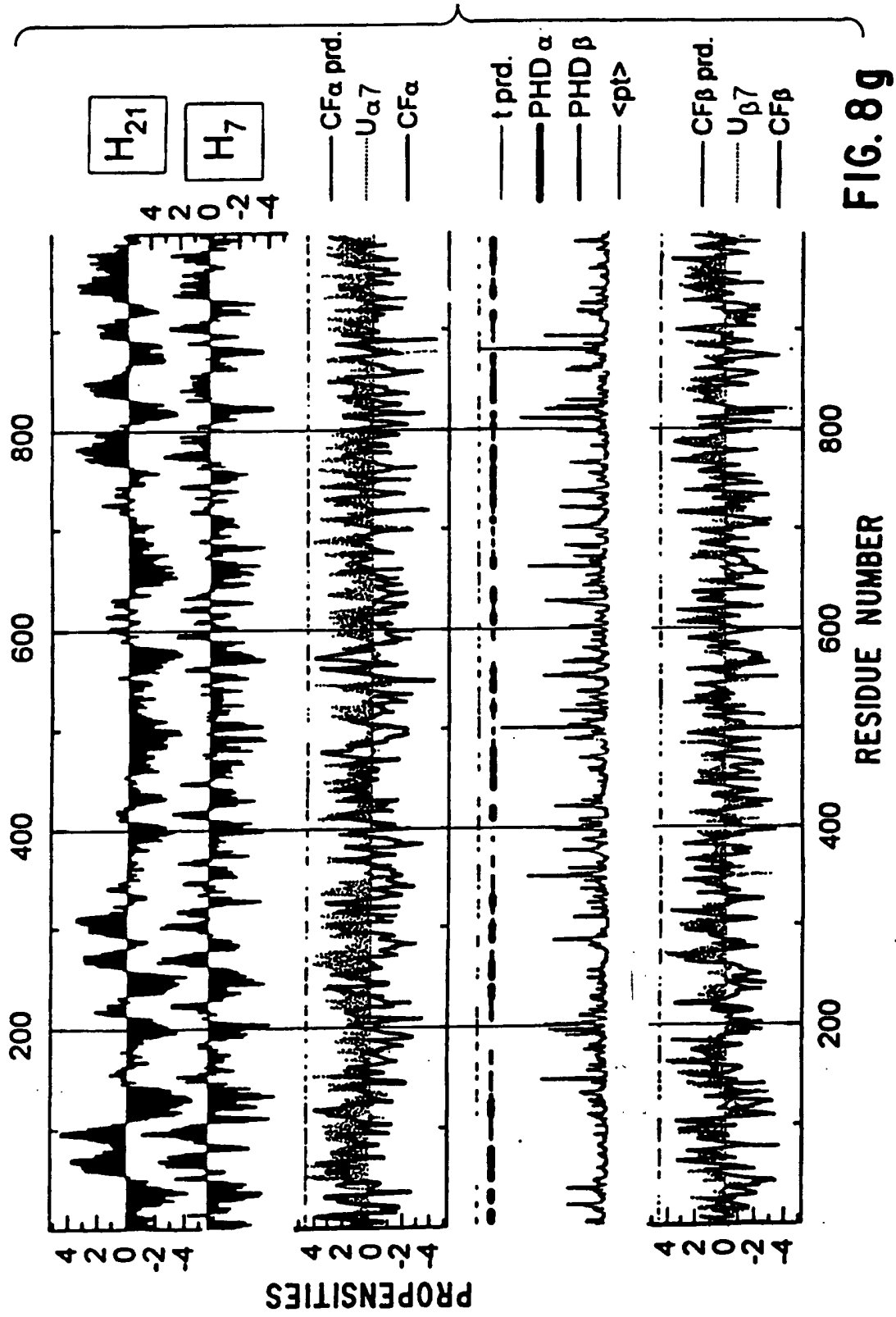

H21
H7
CFα prd.
Uα7
CFα
t prd.
PHD α
PHD β
<pt>
CFβ prd.
Uβ7
CFβ
4 2 0 -2 -4
200
400
600
800
RESIDUE NUMBER
PROPENSITIES

FIG. 8g

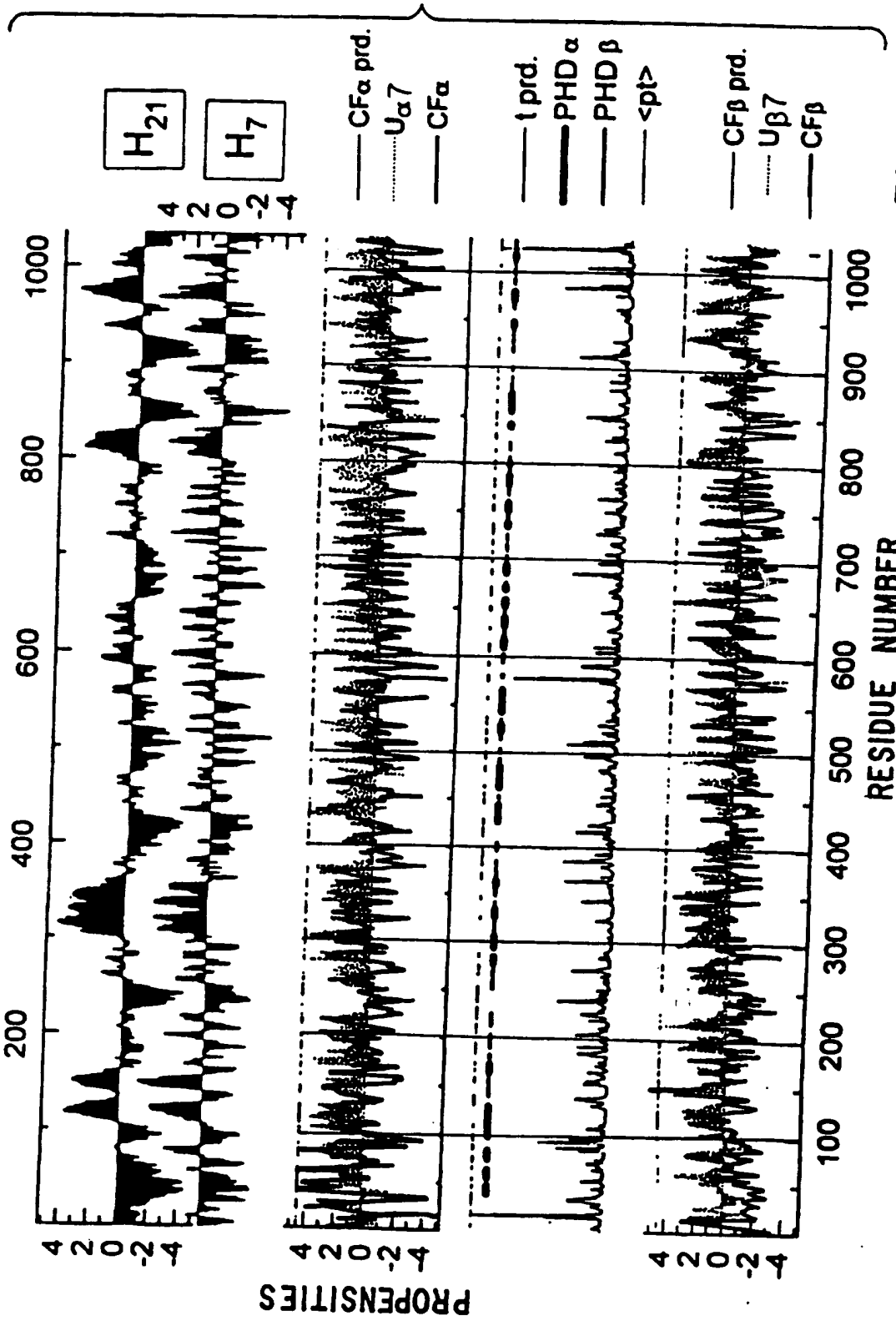

H21
H7
CFα prd.
Uα7
CFα
t prd.
PHD α
PHD β
<pt>
CFβ prd.
Uβ7
CFβ
PROPENSITIES
RESIDUE NUMBER

FIG. 8h

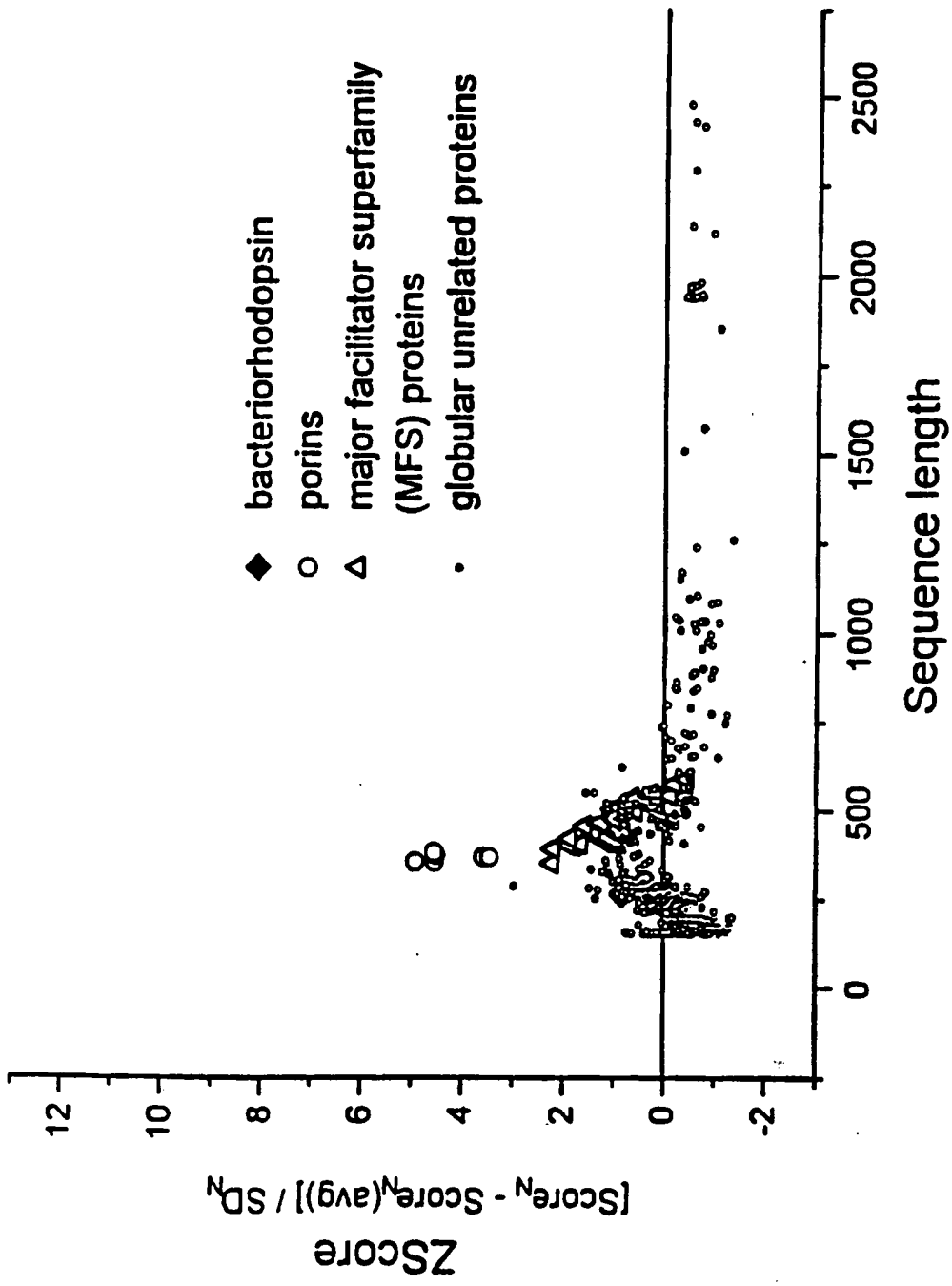

bacteriorhodopsin
porins
major facilitator superfamily (MFS) proteins
globular unrelated proteins
Sequence length
0
500
1000
1500
2000
2500
ZScore
[Score$_N$ - Score$_N$(avg)] / SD$_N$
12
10
8
6
4
2
0
-2

FIG.9a

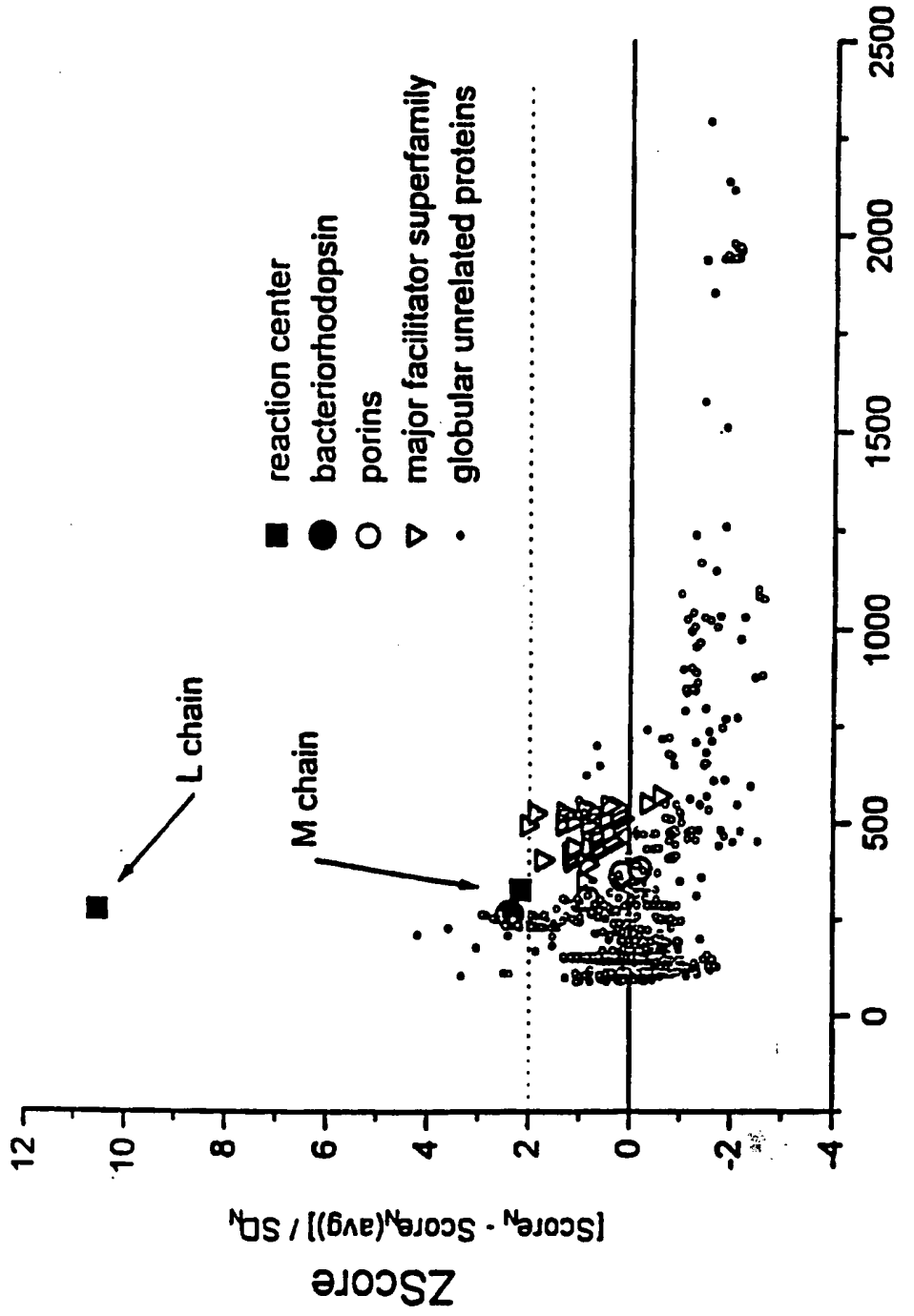

ZScore
[Score$_N$ - Score$_N$(avg)] / SD$_N$
12
10
8
6
4
2
0
-2
-4
0
500
1000
1500
2000
2500
Sequence length
reaction center
bacteriorhodopsin
porins
major facilitator superfamily
globular unrelated proteins
L chain
M chain

FIG.9b

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US95/16126

A. CLASSIFICATI N F SUBJECT MATTER

IPC(6) :G06F 17/10, 17/50, 19/00
US CL :364/496, 578
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 364/496-499, 578

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, PROQUEST, DIALOG, STN

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	PROC. OF NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA, VOL. 90, NO. 24, ISSUED 15 DECEMBER 1993, FISCHBARG ET AL, "EVIDENCE THAT FACILITATIVE GLUCOSE TRANSPORTERS MAY FOLD AS [BETA]-BARRELS", PP. 11658-11662, ESPECIALLY THE METHODS SECTION.	1, 4-10, 13-18.
A	US, A, 5,265,030 (SKOLNICK ET AL) 23 NOVEMBER 1993, ALL SECTIONS.	1-18.
A	US, A, 4,853,871 (PANTOLIANO ET AL) 01 AUGUST 1989, ALL SECTIONS.	1-18.
A	US, A, 4,939,666 (HARDMAN) 03 JULY 1990, ALL SECTIONS.	1-18.

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be part of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search 02 APRIL 1996	Date f mailing f the international search report 22 APR 1996
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized fficer EMANNUEL T. VOELTZ Telephone No. (703) 305-9714

Form PCT/ISA/210 (second sheet)(July 1992)*

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US95/16126

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, f the relevant passages	Relevant to claim No.
A	BIOTECHNIQUES, VOL. 14, NO. 6, ISSUED 1993, HOLBROOK, SR. ET AL, "PROBE: A COMPUTER PROGRAM EMPLYING AN INTEGRATED NEURAL NETWORK", PP. 984-989.	1-18.
A	SYSTEM SCIENCES, 1994 ANNUAL HAWAII INT'L CONFERENCE, VOL. 5, ISSUED 1994, MILLER ET AL, "IDENTIFYING REPEATED STRUCTURAL ELEMENTS IN FOLDED PROTEINS", PP. 235-244.	1-18.
A	INDUSTRIAL FUZZY CONTROL AND INTELLIGENT SYSTEMS, 1993 INT'L CONFERENCE, ISSUED 1993, DAUGHERITY, "A NEURAL-FUZZY SYSTEM FOR THE PROTEIN FOLDING PROBLEM", PP. 47-49.	1-18.
A	BIOCHEMISTRY, VOL. 31, NO. 26, ISSUED 1992, YOSHIMURA ER AL, "FUSION OF PHOSPHOLIPID VESICLES INDUCED BY AN AMPHIPHILIC MODEL PEPTIDE: CLOSE CORRELATION BETWEEN FUSOGENICITY AND HYDROPHOBICITY OF THE PEPTIDE IN AN [ALPHA]-HELIX", PP. 6119-6126.	1-18.
A	SCIENCE, VOL. 223, ISSUED 20 JANUARY 1984, KAISER ET AL, "AMPHIPHILIC SECONDARY STRUCTURE: DESIGN OF PEPTIDE HORMONES", PP. 249-256.	1-18.